Scalable Computing: Practice and Experience

Scientific International Journal for Parallel and Distributed Computing

ISSN: 1895-1767

Universitatea de Vest
din Timişoara

Volume 26(3)

May 2025

Editor-in-Chief

Dana Petcu
West University of Timisoara, Romania

Senior Editor

Marcin Paprzycki
Systems Research Institute of the Polish Academy of Sciences, Poland

Executive Editor

Katarzyna Wasielewska-Michniewska
Systems Research Institute of the Polish Academy of Sciences, Poland

Technical Editor

Silviu Panica
Institute e-Austria Timisoara, Romania

Editorial Board

Peter Arbenz, Swiss Federal Institute of Technology,
Giacomo Cabri, University of Modena and Reggio Emilia,
Philip Church, Deakin University,
Frederic Desprez, INRIA Grenoble Rhône-Alpes and LIG laboratory,
Yakov Fet, Novosibirsk Computing Center,
Giancarlo Fortino, University of Calabria,
Gianluca Frasca-Caccia, University of Salerno,
Fernando Gonzalez, Florida Gulf Coast University,
Dalvan Griebler, Pontifical Catholic University of Rio Grande do Sul,
Frederic Loulergue, University of Orleans,
Svetozar Margenov, Institute for Parallel Processing and Bulgarian Academy of Science,
Fabrizio Marozzo, University of Calabria,
Gabriele Mencagli, University of Pisa,
Viorel Negru, West University of Timisoara,
Wiesław Pawłowski, University of Gdańsk,
Shahram Rahimi, Missisippi State University,
Wilson Rivera-Gallego, University of Puerto Rico,

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

Scalable Computing: Practice and Experience

Volume 26, Number 3, May 2025

TABLE OF CONTENTS

Papers in the special issue on Internet of Things and Autonomous Unmanned Aerial Vehicle Technologies for Smart Agriculture Research and Practice:

Papers in the special issue on Efficient Scalable Computing based on IoT and Cloud Computing:

Papers in the special issue on Adaptive AI-ML Technique for 6G/ Emerging Wireless Networks:

Papers in the special issue on Unleashing the power of Edge AI for Scalable Image and Video Processing:

Papers in the special issue on High-performance Computing Algorithms for Material Sciences:

Papers in the special issue on Deep Learning in Healthcare:

Papers in the special issue on Disruptive IoT-enabled Wearable Systems for Scalable Computing Technologies:

Papers in the special issue on Transformative Horizons: The Role of AI and Computers in Shaping Future Trends of Education:

Papers in the special issue on Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments:

© SCPE, Timișoara 2025

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 9–16, DOI 10.12694/scpe.v26i3.3980

DEVELOPING A FRAMEWORK FOR DETECTING PLANTATION-ROWS THROUGH UAV IMAGE DATA

NISHA M. SHRIRAO[*], VINAY KUMAR SADOLALU BOREGOWDA[†], JAGMEET SOHAL[‡], ANJALI SINGH[§], SACHIN S. PUND[¶], AND RISHABH BHARDWAJ[||]

Abstract. With the introduction of cutting-edge technologies, especially Unmanned Aerial Vehicles (UAVs) or drones, agricultural surveillance and management have seen tremendous changes. These aerial platforms have the ability to obtain high-resolution images across extensive agricultural regions, which can provide crucial information about the health of crops, patterns of growth, and field conditions. Precision agriculture relies on precise identification and monitoring of plantation rows for crop yield estimation, resource allocation, and farm management, requiring labor-intensive physical examination or satellite imagery. The aim of the articles is to create an entire framework that uses a deep learning (DL) approach called the Hungarian optimized fully convolutional deep neural networks (HO-FCDNN) approach, advanced image processing methods, and spatial analysis to identify plantation rows in UAV image data. The UAV image data was first pre-processed using gray-scale Transformation techniques such as feature extraction employing Scale Invariant Feature Transform (SIFT) and image segmentation using vegetation index. Then, using the collected features, novel HO-FCDNN was used to create prediction models that could identify plantation rows. Through the integration of UAV technology with state-of-the-art computational methodologies, the proposed HO-FCDNN technique improved more significantly with the matrices like recall (98.9%), MAE (1.03), F1-score (96.7%) and precision (97.2%). This research aims to enhance sustainable farming practices, improve precision agriculture, and facilitate a more effective use of resources in agricultural production.

Key words: Hungarian optimized fully convolutional deep neural networks, Unmanned Aerial Vehicles (UAVs), plantation rows, agriculture

1. Introduction.

1.1. Unmanned Aerial Vehicles (UAVs) in Plantation Row Detection.

The UAVs or drones made incredible progress in the 20 years at past in the creation of an affordable platform and imaging sensors [3]. Using UAVs is a currently developed, reliable, and growing fine-scale remote sensing technique in precision agriculture that offers several advantages over older field-based and space-borne methods. To capture very-high-resolution (VHR) images in the optical, infrared, and thermal parts of the electromagnetic spectrum, the UAVs are flown with previously unheard-of spatiotemporal resolution [4]. In contrast to human operators and ground-based systems, a UAV was able to fly at any selected time and date without being constrained by ground conditions or clouds. Utilizing advanced remote sensing technology, like light detection and range (LiDAR) has demonstrated outstanding developments in remote sensing technology, as well as wide-area coverage to produce three-dimensional point clouds of agriculture and forests that are very accurate and of the highest quality, facilitating the precision agriculture [5]. Figure 1.1 illustrates the advancements in plantation-row detection.

[*] Electrical Engineering, Yeshwantrao Chavan College of Engineering,Nagpur, Maharashtra, India (1175.nss@gmail.com)

[†] Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Jain (Deemed-to-be University), Bangalore, Karnataka, India (sb.vinaykumar@jainuniversity.ac.in)

[‡] Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh-174103, India (jagmeet.sohal.orp@chitkara.edu.in)

[§] Maharishi School of Engineering & Technology, Maharishi University of Information Technology, Uttar Pradesh, India (anjali.ds@muit.in)

[¶] Department of Mechanical Engineering Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India (pundss@rknec.edu)

[||] Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India (rishabh.bhardwaj.orp@chitkara.edu.in)

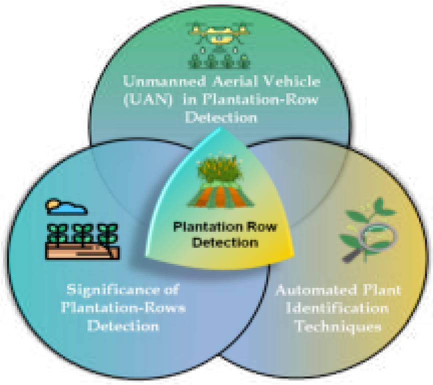

Unmanned Aerial Vehicle (UAN) in Plantation-Row Detection
Plantation Row Detection
Significance of Plantation-Rows Detection
Automated Plant Identification Techniques

Fig. 1.1: Plantation-row detection

1.2. Significance of Plantation-Rows Detection. Row plantation was a common preparation in gardening systems or conventional horizontal farming. The system was composed of several plantations organized linearly in one direction, often east-west. For optimal light exposure, optimum production enhancement, and inventory ease, parallel equidistant rows and uniform seedling spacing are preserved. For the majority of plants that are transplanted, directly seeded, or developed through vegetative or dormant sprouts, row planting was employed [20]. Row planting was beneficial for precision agriculture, as it facilitates the optimization of crop yield while introducing cutting-edge and economical technology and techniques to streamline crop inventory and management procedures. These tools include global navigation satellites system (GNSS), visualization and scanning sensors and geographic information systems (GIS). The automated monitoring of farms and forests requires the precise location of trees to enable associated activities including plant counting, yield estimation, autonomous vehicle navigation, and site-specific tasks. These methods can enhance precision farming applications, leading to enhanced agricultural system management [10].

It is essential to advance agricultural precision by developing a system for detecting plantation rows employing UAV image data. This model improves crop monitoring and reliable management of the field by utilizing high-resolution UAV images and DL methods known as HO-FCDNN, it is beneficial by providing reliable data on crop health and development technologies. It enhances the allocation of research, reduces expensive field control requirements, and facilitates sustainable farming methods.

1.3. Automated Plant Identification Techniques. In general, there are two types of automated plant identification: detection and delineation. Plant size and the image's spatial resolution might be regarded as sufficient characteristics for detection. However, delineation can need knowledge of the objects' spectral heterogeneity, the intricacy of the shadows, and background factors (such as soil brightness). Initially, plants in both cultivated and wooded regions were counted using morphological procedures and algorithms for segmentation such as Valley Following, Region Growing, and Watershed [9]. These techniques primarily depend on the harmonic dispersion across the pixels (non-crop and crop), which shows that the border of the plant is represented by dark pixels, which appear as shadows, and brighter pixels are identified as the plant.

1.4. Objective of this Research. To effectively detect the plantation rows from UAV image data, the purpose of this study is to create a model called HO-FCDNN which is an efficient image processing and spatial analysis technique. By enhancing crop condition detection, resource allocation, and general farm management, this method intends to improve precise agriculture and advance sustainable farming methods. The HO-FCDNN framework can enable in precise identification of plantation rows in agricultural areas, improving precision agriculture techniques and optimizing resource use for sustainable farming.

Structure of the paper. The literature survey is presented in section 2. Section 3 explored the materials and methods. Results and discussions are represented in Section 4. Section 5 is showing the conclusions.

2. Literature Survey. In sugarcane fields, the article [16] used digital image processing on aerial data to automate crop row recognition and spacing measurements. As compared to manual approaches, the results demonstrate a 1.65% relative inaccuracy. Graph-based DL strategy [8] was used to improve plantation determination of lines in UAV-based Red Green Blue (RGB) imagery. It achieved superior performance with 98.7% precision, 91.9% recall and 95.1% F1-score. However, it needed to address generalizability and performance in various environments. Using computer methods and aerial imagery, they attempted to automate spacing measuring and crop row identification in Brazilian sugarcane fields [15]. The results demonstrated the possibility of efficient sugarcane plantation monitoring, with an error rate of less than 2% when compared to manual measurements. By utilizing digital computer vision, image processing, and machine learning (ML) [6], sought to automate crop row recognition and spacing measuring in sugarcane fields. The findings indicate a low mapping error of 1.65% when compared to human mapping; nevertheless, there could be restrictions on how broadly the results can be applied in different field situations. They used DL to automate citrus orchards [18] surveillance through gap identification, tree detection, and seedling identification using drone-captured aerial images. Reaching great accuracy for fully grown trees, problems with gap and seedling recognition rather exist and require more improvement. The article proposed a method to mitigate the high cost of human data labeling in agriculture by improving plant row and spacing detection with the use of SegFormer and domain adaption [7]. With an emphasis on ground-level camera placement [2], proposed to establish a direct strategy of employing neural networks to recognize crop lines from RGB images in fields of crops. Findings indicate precise row identification and location prediction; however, performance could be impacted by environmental variability [1]. Paper [12] offered a convolutional neural network (CNN) method to address the challenge of assessing the number of citrus trees in highly dense orchards from UAV multispectral images. The method outperforms significantly object detection methods for counting and geolocation. The goal of the study [17] was to use a novel approach to measure inequalities in crop fields and identify crop rows. Utilizing a tiny remotely piloted aircraft, the mosaic of actual scene images was used to assess the suggested approach. In the areas where there were absences in the curving crop rows, the planting regions' relative error in the experimental tests was lower than that of manual mapping. Using geometric descriptor data from remote sensing data collected by UAVs, the study [13] established and deployed a system that used innovative DNN and initial maize plant count using machine vision algorithms and finding emphasizes that have minimal emergence in the fields. It was not constrained by the landscape and to achieve reliable performance, it could automatically modify its parameters based on the UAV's flying position. Study [11] compared the three deep learning techniques (DL) for identification of sorghum head with RGB UAV imagery EfficientDet, SSD, and YOLOv4. Following the model parameters analyzed, it was found that the overlapping ratios, confidence, and IoU produced the most effective results for sorghum head detection accuracy. Research [19] suggested an approach that estimates the perspective and uses positions of tree branches. The experimental analysis demonstrated that the suggested method was more effective than the state-of-the-art. Study [14] categorized the images using the CNN and identified the locations in crop rows or unplanted soil. The suggested task's efficiency and viability were demonstrated by the systems' outcomes.

2.1. Research Gap. Through the application of digital processing, DL techniques, and UAV-based imaging, the research on detecting plantation rows has advanced significantly, obtaining high accuracy and precision across a variety of crops. There are some research gaps, mainly related to the robustness and generalizability of the existing articles in a variety of crop types and climatic conditions. The existing approach's versatility is limited by its reliance on certain data sources like UAVs and ground-level cameras. More enhancements are required to make it more effective in managing the field of challenging conditions such as modifying row curvature, imperfections, and seedling recognition. To utilize those techniques for more extensive agricultural applications, it is necessary to address the high cost of data labeling. Through the integration of advanced DL algorithms with UAV images, the proposed HO-FCDNN approach resolves the above issues. It decreases the cost of data labeling, effectively manages complicated environments, and improves generalizability and durability in a variety of conditions. This proposed method advances precise performance in plantation row detection and significantly increases the parameter's performance.

3. Materials and Methods. In this section, the study areas, pre-processing by grayscale transformation, image segmentation, SIFT technique for feature extraction, and classification for detecting plantation

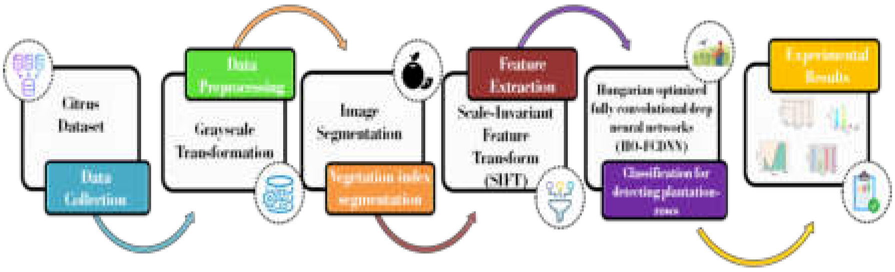

Citrus Dataset
Data Collection
Data Preprocessing
Grayscale Transformation
Image Segmentation
Vegetation index segmentation
Feature Extraction
Scale-Invariant Feature Transform (SIFT)
Hungarian optimized fully convolutional deep neural networks (HO-FCDNN)
Classification for detecting plantation-rows
Experimental Results

Fig. 3.1: Proposed Methodology

rows utilizing an innovative HO-FCDNN technique development is explored. Figure 3.1 depicts the proposed methodology.

3.1. Study Area. The article was performed utilizing citrus orchard (Citrus Sinensis Pera) trees. The experiment was performed on a Citrus Sinensis Pera orchard in a rural region. Citrus trees in the mature stage make up the area and because of the planting techniques, the trees are spaced differently. Later years allowed the planting of more current trees in a more accurate area, resulting in varied densities and spacing in line from each other than the original 3 meters. The area is around ten thousand square meters. Utilizing an $X7$-Spire II UAV equipped with an RGB sensor at an altitude of 80 meters during flight.

3.2. Image Pre-processing using Grayscale Transformation. Pre-processing is initially utilized to distinguish green pixels (crops) from the remainder of the images (soil, stones, and other unexpected things) utilizing fundamental image processing operations. Greyscale images were used for the analysis of images to lower the computing cost. Techniques for removing image noise are needed since all the initial citrus images of the cotton areas such as dirt, gravel, weeds, and other unwanted noise. Image noise could be decreased with the right greyscale transformation. A $2G-R-B$ index method was proposed, which can protect various vegetation types from others in various kinds of natural lighting situations, preventing segmentation. Generally, the primary color element of green plants is green(G), while the color space of the gravels, soil surface, and other background primarily consists of red (R) and blue (B) components. However, certain background areas also exhibit pronounced G components that are insensitive to attenuation using the traditional $2G-R-B$ index. If the G element is superior to the R and B elements, the color pixel P_0's greyscale is determined using the $2G-R-B$ index. If not, the greyscale of P_0 will be allocated a zero value. Thus, Equation (3.1) can be used to generate the greyscale $f(i,j)$ of P_0:

$$e(a,b)=\begin{cases}2G(a,b)-R(a,b)-B(a,b), & G(a,b)>B(a,b)>R(a,b)\end{cases} \quad (1)$$

where $(j,\ i)$ denotes the pixel coordinates of P_0 and $R\,(a,\ b)$, $G\,(a,b)$ and $B\,(a,\ b)$ denote the corresponding red, green, and blue chromatic parameters for P_0. It is possible to minimize background noise and distinguish green plants from the backdrop more successfully using the enhanced $2G-R-B$ index approach. After the pre-processing stage, citrus image data are modified for additional analysis and recognition of features through the segmentation phase, which improves the process's accuracy and efficiency.

3.3. Image Segmentation. Vegetation index segmentation is used in the pre-processed images to separate and differentiate vegetation regions, hence improving the study of vegetation patterns and attributes within the images. Segmentation process as much as feasible is important for real-time applications. Utilized a learning-based approach, our objective was to determine the percentage of the green spectral component compared to the rest, allows to classify a pixel as part of a green plant. This strategy contrasts with using vegetation indices, which necessitate an image transforming a UAV image from RGB color space to gray-scale. To determine differences among the 3 spectrum RGB elements that distinguish green plants, this relative percentage is meant to deal with light fluctuation. This is accomplished by using a fuzzy clustering technique. This method consists of two phases: a period of education where the threshold or relative percentage is calculated offline during an activity and a choice stage at which the threshold is used without further calculation.

The structure of the learning phase proceeded as outlined below. First randomly selected m samples ($W = \{w_1, w_2, ..., w_n\} \in Q^c$, where c is the dimensionality of the data) from the collection of accessible images. The three RGB spectral components of each image's pixel at the initial citrus image position $(w,\ z)$ make up each sample vector(w_j). This indicates that the data dimensionality for these trials is $c = 3$. Every sample must be allocated to a specific cluster, denoted as w_i, out of a total of c potential clusters, or $i = 1,\ 2, ..., d$. Since were only focused on two sorts of textures in natural plants (crops/weeds) and the break (detritus, stones and soil) d are set to 2 in the suggested technique. The common fuzzy clustering technique assumes that the number of clusters d is known to be utilized for categorizing the samples in X. It utilizes as input the samples w_j and creates to allocate. The procedure maintains the cluster centers u_i and calculates each w_j's degree of membership in the cluster $\mu_j^i\,(s+1)$ at iteration s shown in equations (3.2) and (3.3).

$$\mu_j^i\,(s+1) = \frac{1}{\sum_{q=1}^{d} \left(c_{ji}(s)/c_{jq}(q)\right)^{2/(a-1)}} \tag{2}$$

$$u_i\,(s+1) = \frac{\sum_{j=1}^{m} \left[\mu_j^i\,(s)\right]^a w_j}{\sum_{j=1}^{m} \left[\mu_j^i\,(s)\right]^a} \tag{3}$$

Euclidean distance squared is $c_{ji}^2 = c^2(w_j,\ u_i)$. The exponential weight, $a > 1$, is represented by the integer a. When either a number tmax of iterations is reached or $-$for every ji, the iteration process stops meeting its stopping requirement. To execute equation (3.2) at iteration $S = 1$, the procedure necessitates initializing the cluster centers. To utilize the pseudo-random process mentioned as follows:

- The training sample values should be transformed linearly ($Z = e\ (w)$) such that they drop inside the interval [0, 1].
- Set up $u = 2C\bar{N} \circ Q + C\bar{n}$, where m represents the mean vector of the transformed training samples into Z and $\bar{n} = max(abs(Z - \bar{n}))$, both of $size$ $1 \times c$; $C = [1...\ 1]^S$ of $size$ $d \times 1$; Q is a $d \times c$measures of arbitrary numbers in $[0,\ 1]$, the operation $\circ$ indicates the multiplication of individual elements.

Following learning, clusters x_1 and x_2 have two cluster centers, u_1 and u_2, while u_1 denotes green plants. The equation for differentiating between green and non-green plants is $S_H = u_{1H}/(u_{1Q}+u_{1H}+u_{1A})$. Assuming the matching RGB pixels have an H spectral value larger than S_H, the green areas of the images are recognized once S_H is obtained. Transferring information to feature the extraction procedure, where pertinent features are found and retrieved for additional evaluation and decision-making, is a step in the image segmentation process.

3.4. Extrating Features by Employing Scale-Invariant Feature Transform (SIFT). The difference of Gaussians is employed by the SIFT technique to determine the difference of Gaussian (DOG) scale-space. The process involves extracting feature points from citrus plantation images with varying spatial scales, eliminating unstable feature points, determining the primary direction of the characteristic points, and generating a SIFT characteristic descriptor to prevent mismatches due to noise, rotation, illumination, and scale of plants. Scale-space's fundamental concept is to gather visual processing data at various sizes and use in-depth analysis to identify the key components of these images. To determine the scale-space $K(w,\ z,\ \sigma)$ of plant image $k(w,\ z)$, which is shortened as K, by integrating with the Gaussian functions $H(w,\ z,\ \sigma)$ of scale component σ. Using scale-space $K(w,\ z,\ \sigma)$, SIFT further develops the DOG scale-space $C(w,\ z,\ \sigma)$, which is shortened as C, to increase the effectiveness of feature point extraction. Each pixel of crop images is compared with eight neighboring pixels at the same scale and nine surrounding pixels of crop images at the two greatest levels in the DOG scale-space C of crop image $k(w,\ z)$. This pixel is regarded as a feature point if it is the minimum or maximum value. The low-contrast feature points must be deleted since the locations of the feature points acquired in this manner are offset. To determine the location of feature point B' with more accuracy, the Taylor series expansion of the scale-space functions C at point B (Equation (3.5)).

$$C = C_B + \frac{\partial C_C^S}{\partial w}\, w + \frac{1}{2} w^S \frac{\partial^2 C_B}{\partial_w{}^2}\, w \tag{4}$$

Here the offset via point B to point B' is given by $w = (w, z, \sigma)T$. To determine the exact position of feature point B' in (Equation (3.6)), is to compute the derivative on both sides of $C(w)$.

$$\widehat{w} = \frac{\partial^2 \partial_B^{-1}}{\partial_w{}^2} \frac{\partial C_B}{\partial w} \tag{5}$$

To compute the DOG scale-space function $C(\widehat{w})$ of B' (Equation (3.7)).

$$C\left(\widehat{w}\right) = C_B + \frac{1}{2} \frac{\partial C_B^S}{\partial w} \widehat{w} \tag{6}$$

Dc is employed to test the reliability of crop feature points and it is set to the contrast threshold $|C(\widehat{w})|$ in the DOG scale-space. Typically, C_d is set to 0.03; feature points with C_d less than 0.03 are eliminated. However, several investigators observed that the standard C_d could not be optimal for every cropped image, leading to reduced effectiveness and precision in the feature point extraction process. Therefore, the developers must choose the proper contrast threshold in the DOG scale space for various UAV image types utilized for plantation row detection. When characteristics are extracted from citrus plantation images, the HO-FCDNN approach is employed to improve the precision and effectiveness of plantation row identification.

3.5. Detecting Plantation-Rows using Hungarian optimized fully convolutional deep neural networks (HO-FCDNN).

3.5.1. Fully Convolutional Deep Neural Networks (FCDNN). A type of feed-forward neural network known as FCDNNs is the neural network's biggest category. Vector data is fed into an FCDNN, which produces another vector. An FCDNN is composed of many completely linked layers, with numerous nodes in each fully connected layer. The input layer nodes allow crop data to reach the FCDNN. Every node has connections to every other node in the previous layer. All node elements are the weighted average of the nodes' elements through the layer before it. The parameters in FCDNN that can be trained are called weights. Though the output layer is usually linear, the hidden layer nodes' outputs usually pass via exponential linear units are the representation of a function of non-linear activation. In general, a plantation row predicted amount is represented by the number of every output layer node. For this reason, numerous numbers can be plantation rows predicted concurrently using FCDNNs. DNN consists of a subset of fully linked networks. These completely linked networks are referred to as "Structure Agnostic" networks. The 6-layer network framework of the FCDNN consisted of the input layer, 4 hidden ones, and the output layer in the research, each of which was made up of many neurons that could be computed concurrently. An activation function linked the hidden layers together, as well as the first hidden layer and the input layer. Figure 3.2 depicts the structure of FCDNN. Following are the specifics of the suggested FCDNN's structure.

Fully connected layers or dense layer. A significant degree of efficacy in learning non-linear mixes of input crop attributes was shown by the completely linked layers. The neurons in a completely connected layer are linked with activation in the layer that came before it. The matrix multiplication and bias offset in equation (3.8) can be used to determine their activations.

$$G(w) = Xw + a, \tag{7}$$

Here the biased offset is $a \in Q^L$ and the weight measure is $X \in Q(L, m)$.

Exponential Linear Unit (ELU) Activation Layer. The ELU is a function that typically converges more quickly and accurately in equation (3.9).

$$ELU(w) = \left\{ \begin{array}{c} \alpha\left(f^w - 1\right) \text{ if } w \ \leq 0 \\ w \text{ if } w > 0 \end{array} \right\} \tag{8}$$

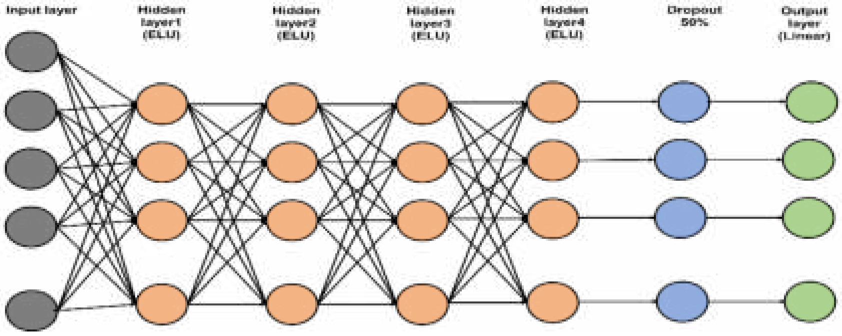


Fig. 3.2: Structure of FCDNN

Since the learning process can be increased, ELU leverages the activation function to arrive at mean zero. An α value is selected for the function of ELU activation; a typical value falls between 0.1 and 0.3. Because it reduces prejudice change obtained by reducing the mean activation lower to zero, it is a useful alternative function of activation such as ReLU (Rectified Linear Unit).

Dropout layer. A kind of regularization known as dropout involves randomly removing a certain percentage of the link streaming into a completely connected layer. When a connection is dropped, its influence on the associated function of activation is set to 0, which stops the system from learning value by excessively fitting. Even for large, deep networks, dropout will cause training loss to stop rapidly approaching zero.

Linear activation layer. Equation (3.10) describes the linear activation function.

$$B = dw \tag{9}$$

Where crop input controls activation and d is a constant quantity. This makes it a non-binary activation as it offers a variety of activations. The plant row detection process with FCDNN, the Hungarian-optimized technique will be employed to improve and optimize the citrus plantation data.

3.5.2. Hungarian Optimization. The Hungarian method improves the relationship between the centroids of the actual target positions and the centroids of the anticipated trajectories in the crop rows. The Euclidean distance from the targets' actual crop locations and the center points of the projected trajectories is the cost function. The task of assigning becomes the crop row challenge of determining a bipartite graph's best solution.

Phase 1. Determine the quantity that it will cost to map the targets' crop locations onto the platforms. The centroids of the anticipated tracks and the crop locations of the actual target that was found are divided into Euclidean distances. The equation appears in Equation (3.11).

$$c = \sqrt{\textstyle\sum_{j=1}^{M} (q_{1j} - q_{2j})^2} \tag{10}$$

Here the anticipated trajectories and the centroids of the actual target placements are denoted by $q_{1j}$and q_{2j}, respectively. The Euclidean distance is expressed by the variable c and all computed distance values are retained.

Phase 2. The Hungarian method was employed to maximize the crop allocation of estimated trajectories and real target centroids. The method has been modified to identify the best solution for a bipartite graph using a $N\ x\ 2$ measure. For an accurate plantation row forecast of target positions, the Hungarian approach makes several assumptions, such as a maximum of M target positions and one trajectory allocated to each target, as demonstrated in Equation (3.12).

$$L = (W,\ Z,\ F,\ X) \tag{11}$$

Here $L.\ x$ denotes the weight of each edge in the bipartite graph, $L.\ x$ is a collection of all the locations of the actual targets, and F denotes the set of every edge in the partial graph. A collection of all the anticipated tracks' places is represented by Z. The crop row values of sets W and Z are given by equations (3.13) and (3.14).

$$W = \{w_1, w_2, \ldots, w_N\} \tag{12}$$

$$Z = \{z_1, z_2, \ldots, z_m\} \tag{13}$$

The label of a vertex in the bipartite graph, L, is thought of as mapping across the set of positive integers and the set of edges in the bipartite graph. Equation (15) displays the weights of the edges in L.

$$\begin{array}{ccccc} X & Z_1 & Z_2 & \cdots & Z_M \\ W_1 & X_{11} & X_{12} & \cdots & X_{1M} \\ W_2 & X_{21} & X_{22} & \cdots & X_{2M} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ W_N & X_{N1} & X_{N2} & \cdots & X_{NM} \end{array} \tag{14}$$

The label of a vertex in the weighted bipartite graph L is referred to as an executable vertex when the weight (x) of any edge is less than the minimum weight (T) of any side of the present vertex.

To locate the ideal match for assigning the crop locations of the trajectories and targets, one must locate the equal sub-graph, or L, of the bipartite graph. In the bipartite graph with weights, L is in equation (3.16).

$$F_J = \{wz\ \in F\,(w)\,|U\,(w) + U\,(z) = X\,(wz)\tag{15}$$

Here El is the equal subgraph of the bipartite graph L and $F(L)$ is a collection of all the edges in the partial graph G. $U\,(z)$ represents the label of vertex W's positive crop integer. $U\,(z)$ is the label of the positive crop integer of edge z. Then, it is possible to consider the subgraph El that was created in L as an identical subgraph of L. The crop locations of the trajectories and targets can be optimally allocated, and the non-assigned trajectories will be eliminated after a particular number of frames in plantation row detection. An effective system for identifying plantation rows in agricultural images is the HO-FCDNN. DL is used to extract complex characteristics and HO ensures sure that pixels are assigned accurately, improving efficiency for applications like crop tracking and yield estimation. The HO-FCDNN design is a dependable and effective precision agricultural solution as it is resistant to changes in weather, illumination, and crop varieties.

4. Result and Discussion. The gathered crop images were divided into 564 non-overlapping patches, each measuring 256 × 256 pixels. The images of the citrus orchards were divided into 635 identical patches, each measuring 256×256 pixels. Additionally, the line and point characteristics that were recognized as image samples were citrus regions. 235 citrus plantation rows and 16925 trees were utilized in the study to create the citrus orchard. Determining the generality and robustness of the HO-FCDNN technique required this kind of testing using various plant phonologies, locales, and sensor features.

4.1. Experimental Setup. The HO-FCDNN techniques were implemented in an Intel $i3 - 4330TE$ dual-core CPU running at 120 GB, 4 GB of RAM, and a 2.4 GHz hard drive are the specifications of Jackal's computer. An external personal computer linked to the Jackal through Ethernet processes the UAV images, while the Jackal uses a Robot Operating System (ROS) to control the device's mobility.

4.2. Plantation Row Evaluation in Citrus Dataset. The variables for the citrus plantation are identical to those selected for the citrus plans ($\sigma_{max,\ plant} = 3$, $\sigma_{min,row} = 0.5$, and $\sigma_{max,\ row} = 3$, and $S\ =\ 6$). The only difference is the maximum pixel distance (15 pixels for the citrus radius, compared to 8 pixels in corn), which makes sense given the variation in size between the canopy's areas of the citrus trees. The obtained findings demonstrate that requiring only minor simulation modifications, the HO-FCDNN methodology can be used for many kinds of plantations (Figure 4.1 (a) and (b)). Furthermore, the HO-FCDNN technique continues to function well and yield very accurate forecasts even in high-density plantations like the citrus data sets. This demonstrates that the MSM program is beneficial not only for learning about plant development in many phases but also for learning about different kinds of plantations with more difficult high-density.

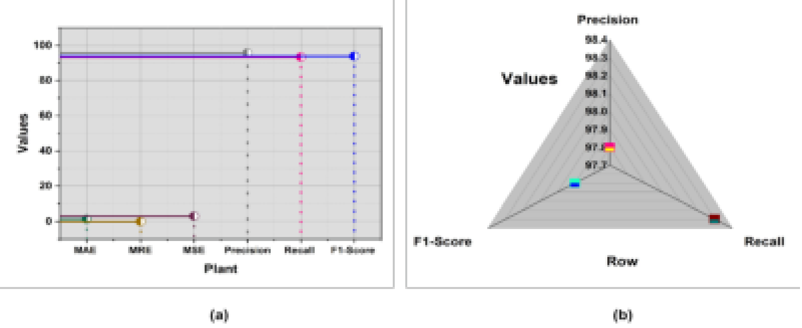


Fig. 4.1: (a) and (b): Outcome of HO-FCDNN method for citrus orchard dataset

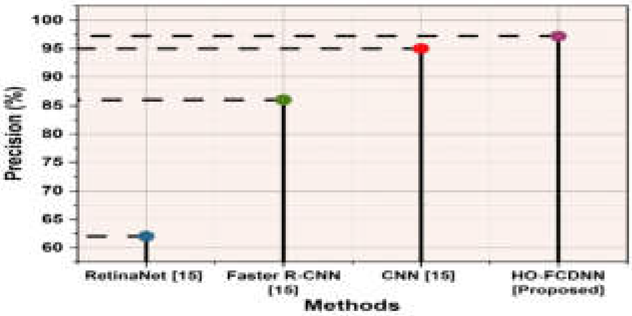


Fig. 4.2: Result of Precision

4.3. Performance Evaluation. This section compares the effectiveness of the recommended HO-FCDNN approach and the traditional method using metrics: the Mean Absolute Error (MAE), precision, recall, and f1-score. Some of the current methods are Conventional Neural Networks (CNN) [12], Faster Region-based Convolutional Neural Networks (Faster R-CNN) [12] and RetinaNet [12].

Precision. Precision is a measure of how accurate a technique is at generating positive predictions. It is calculated as the ratio of all the accurate forecasts to all the inaccurate forecasts. It pertains to plantation row detection, accuracy would be the percentage of projected rows that match the actual number of rows. Figure 4.2 represents the comparison outcome of the suggested method. The HO-FCDNN method is attaining 97.2% a greater result than current methods [12] like RetinaNet is 62%, CNN is 95%, and Faster R-CNN is 86%. The HO-FCDNN approach is very useful for the detection of plantation rows.

Recall. Recall in the setting of plantation row detection would show the proportion of real rows that the model properly detected. The proposed method is compared to current methods, shown in Figure 4.3. The suggested HO-FCDNN strategy performs 98.9% of the greater numerical results compared to an existing method [12] RetinaNet performs 92%, CNN performs 96% and Faster R-CNN performs 39%.

F1 Score. When there is an unequal distribution of classes, it can be beneficial as it takes into consideration fake positives and false negatives. The suggested method is compared to the current approach. The comparison result are showed in Figure 4.4. The HO-FCDNN technique achieves 96.7% and has higher values than existing methods [12] like RetinaNet achieves 74%, CNN achieves 95% and Faster R-CNN achieves 54%, which indicates the HO-FCDNN is a better performance result of detection of plantation row.

Mean Absolute Error (MAE). A model employed in the detection of plantation row is measured by MAE that estimates the total variance average among its initial value and anticipated. The proposed method is compared to current methods, shown in Figure 4.5. Table 4.1 depicts the overall numerical result of the HO-

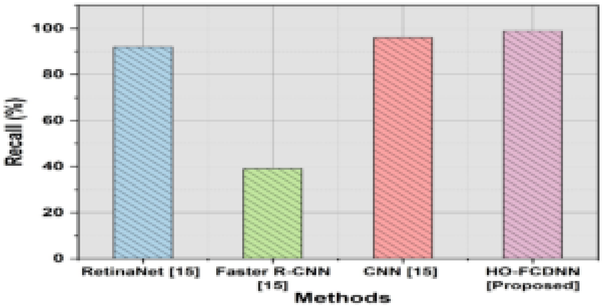


Fig. 4.3: Outcome of Recall

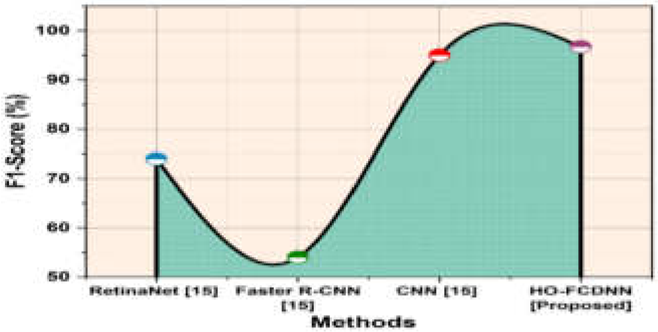


Fig. 4.4: Comparison result of F1-score

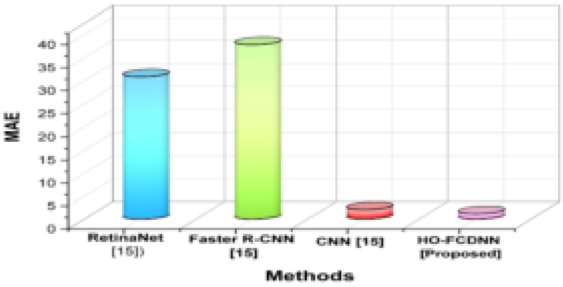


Fig. 4.5: Graphical representation of MAE

FCDNN strategy. The suggested HO-FCDNN strategy attained 1.13 has lower numerical results compared to an existing method like RetinaNet [12] attained 30.87, CNN [12] attained 2.05 and Faster R-CNN [12] attained 37.85. The MOAC-ADenseNet has a significant effect on cost-effective UHPC material selection.

5. Conclusion. The article provided a complete framework that would enable precise identification of plantation rows in UAV images for precision agricultural applications by utilizing ML, sophisticated methods

Table 4.1: Overall Numerical Result of HO-FCDNN Strategy

Methods	**F1-Score (%)**	**Recall (%)**	**Precision (%)**	**MAE**
RetinaNet [12]	74	92	62	30.87
Faster R-CNN [12]	54	39	86	37.85
CNN [12]	95	96	95	2.05
HO-FCDNN [Proposed]	96.7	98.9	97.2	1.13

for image processing, and spatial evaluation. The procedure was pre-processing UAV images using gray-scale transform, by segmenting UAV images employing vegetation indices approach and extracting features using Scale-Invariant Feature Transform (SIFT) and then training newly developed Hungarian optimized fully convolutional deep neural networks (HO-FCDNN) to provide prediction models for plantation row identification. To examine, the efficiency of the HO-FCDNN strategy with a various kind of farming achieved in the citrus plantation dataset. It repaid an MRE of 0.0615 citrus trees per patch, MAE equal to 1.409, F1-score (0.965), recall (0.911), and precision (0.922). For the detection of citrus plantation row, HO-FCDNN strategy the outcome in the recall, F1- scores, and precision are 0.970, 0.964, and 0.965 respectively. For the detection of plantation-row compared to the existing methods, the HO-FCDNN strategy repaid recall (98.9%), precision (97.2%), Mean Absolute Error (1.13), and F1- scores (96.7%) respectively.

6. Limitation and Future Scope. Different environmental factors like varied weather could provide difficulties for the proposed model and have an impact on the reliability of feature extraction and image quality. Its utilization in regions where drone connectivity or data supply exists could be limited by its dependency on high-resolution UAV images. It could be possible in the future to integrate real-time processing for dynamic areas and alter the model to fit a variety of agricultural environments. For further enhancement in row detection and to expand the model's reliability, the algorithm could be improved as more adaptable changes in the environment and include multi-spectral images.

REFERENCES

[1] P. Cinat, S. F. D. Gennaro, A. Berton, and A. Matese, Comparison of unsupervised algorithms for Vineyard Canopy segmentation from UAV multispectral images, Remote Sensing, 11 (2019), pp. 1023–1023.

[2] I. F. D. Costa and W. Caarls, *Crop Row Line Detection with Auxiliary Segmentation Task*, in Brazilian Conference on Intelligent Systems, Springer Nature Switzerland, 2023, pp. 162–175.

[3] V. Czymmek, R. Schramm, and S. Hussmann, *Vision-based crop row detection for low-cost UAV imagery in organic agriculture*, 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), (2020), pp. 1–6.

[4] T. De Swaef, W. H. Maes, J. Aper, J. Baert, M. Cougnon, D. Reheul, and P. Lootens, *Applying RGB-and thermal-based vegetation indices from UAVs for high-throughput field phenotyping of drought tolerance in forage grasses*, Remote Sensing, 13 (2021), pp. 147–147.

[5] S. Debnath, M. Paul, and T. Debnath, *Applications of LiDAR in agriculture and future research directions*, Journal of Imaging, 9 (2023), pp. 57–57.

[6] S. Dhariwal and A. Sharma, *Aerial Images were used to Detect Curved-Crop Rows and Failures in Sugarcane Production*, 2022 IEEE International Conference on Electronics, Computing and Communication Technologies, pp. 1–7.

[7] A. S. Ferreira, J. M. Junior, H. Pistori, F. Melgani, and W. N. Gonçalves, *Unsupervised domain adaptation using transformers for sugarcane rows and gaps detection*, Computers and Electronics in Agriculture, 203 (2022), pp. 107480–107480.

[8] D. N. Gonçalves, M. D. S. D. Arruda, H. Pistori, V. J. M. Fernandes, A. P. M. Ramos, D. E. G. Furuya, and W. N. A. Gonçalves, deep learning approach based on graphs to detect plantation line, (2021), pp. 3213–3213.

[9] J. Gu, H. Grybas, and R. G. Congalton, *Individual tree crown delineation from UAS imagery based on region growing and growth space considerations*, Remote Sensing, 12 (2020), pp. 2363–2363.

[10] H. Kendall, B. Clark, W. Li, S. Jin, G. D. Jones, J. Chen, and L. J. Frewer, Precision agriculture technology adoption: A qualitative study of small-scale commercial "family farms" located in the North China Plain, Precision Agriculture, (2022), pp. 1–33.

[11] H. Li, P. Wang, and C. Huang, *Comparison of deep learning methods for detecting and counting sorghum heads in UAV imagery*, Remote Sensing, 14 (2022), pp. 3143–3143.

[12] L. P. Osco, M. D. S. D. Arruda, J. M. Junior, N. B. Silva, A. P. M. Ramos, E. A. S. Moryia, and W. N. A. Gonçalves, *convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imager*, ISPRS

Journal of Photogrammetry and Remote Sensing, 160 (2020), pp. 97–106.

[13] Y. Pang, Y. Shi, S. Gao, F. Jiang, A. N. Veeranampalayam-Sivakumar, L. Thompson, and C. Liu, *Improved crop row detection with deep neural network for early-season maize stand count in UAV imagery*, Computers and Electronics in Agriculture, 178 (2020), pp. 105766–105766.

[14] J. B. Ribeiro, R. R. Silva, J. D. Dias, M. C. Escarpinati, and A. R. Backes, Automated detection of sugarcane crop lines from UAV images using deep learning, Information Processing in Agriculture, (2023).

[15] B. M. Rocha, A. U. D. Fonseca, H. Pedrini, and F. Soares, *Automatic detection and evaluation of sugarcane planting rows in aerial images*, Information Processing in Agriculture, 10 (2023), pp. 400–415.

[16] B. M. Rocha, G. S. Vieira, A. U. Fonseca, N. M. Sousa, H. Pedrini, and F. Soares, *Detection of Curved Rows and Gaps in Aerial Images of Sugarcane Field Using Image Processing Techniques*, IEEE Canadian Journal of Electrical and Computer Engineering, 45 (2022), pp. 303–310.

[17] ———, *Detection of Curved Rows and Gaps in Aerial Images of Sugarcane Field Using Image Processing Techniques*, IEEE Canadian Journal of Electrical and Computer Engineering, 45 (2022), pp. 303–310.

[18] L. E. C. Rosa, M. Zortea, B. H. Gemignani, D. A. B. Oliveira, and R. Q. Feitosa, *Fcrn-based multi-task learning for automatic citrus tree detection from UAV images*, 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS) IEEE, (2020), pp. 403–408.

[19] R. Silva, J. M. Junior, L. Almeida, D. Gonçalves, P. Zamboni, V. Fernandes, and W. Gonçalves, *Line-based deep learning method for tree branch detection from digital images*, International Journal of Applied Earth Observation and Geoinformation, 110 (2022), pp. 102759–102759.

[20] Z. Yang, Y. Yang, C. Li, Y. Zhou, X. Zhang, Y. Yu, and D. Liu, *Tasseled crop rows detection based on micro-region of interest and logarithmic transformation*, Frontiers in Plant Science, 13 (2022), pp. 916474–916474.

Edited by: Vadivel Ayyasamy
Special issue on: Internet of Things and Autonomous Unmanned Aerial Vehicle Technologies for Smart Agriculture Research and Practice
Received: May 18, 2024
Accepted: Feb 11, 2025

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1005–1016, DOI 10.12694/scpe.v26i3.4190

SMART AGRICULTURE: INTEGRATING AIR QUALITY MONITORING WITH DEEP LEARNING FOR PROCESS OPTIMIZATION[*]

SHOBANA J[†] VENKATA SUBRAMANIAN A[‡] BALAMURUGAN P [§] SIVAKUMAR PERUMAL [¶] SANKARI V[‖] ELDHO K J[**] AND NARESHKUMAR R[††]

Abstract. Modernization and intense industrialization have led to a substantial improvement in people's quality of life. However, the aspiration for achieving an improved quality of life results in environmental contamination. A primary consequence of environmental degradation is air pollution, resulting from rising levels of poisonous chemicals in the atmosphere, which may induce detrimental health conditions in humans. It is harmful to both humans and agriculture. Given that the effects of air pollution on plants may not be readily apparent, it is important to analyse the necessary data and compute the outcomes. Farmers prioritise on pests and plant diseases, frequently neglecting the detrimental impacts of air pollution. Some plant species can withstand high amounts of pollution from suspended particulate matter and accumulated gases, while others are more susceptible to harm. Therefore, plants' reaction to air pollution is influenced by the kind of harmful compounds, their levels, and the plant's susceptibility to them. The LSTM +CNN Proposed Ensemble method may be used to analyse the impact of air pollution on agriculture by examining trends in crop production over time and predicting which crop is more resistant based on the pollution data. The initiative created for this aim may assist farmers in determining the most suitable crop to cultivate in their fields to minimize the impact of air pollution on agricultural yield. The findings show deep learning algorithms correctly predict hourly pollutant concentrations such as carbon monoxide, sulphur dioxide, nitrogen dioxide, ground-level ozone, and particulate matter 2.5, along with the hourly Air Quality Index (AQI) for California. A proposed model used test RMSE values as a measure to evaluate prediction performance, achieving the best possible results.

Key words: Air quality, Deep Learning, Pollution Environmental, neural networks.

1. Introduction. The effects of air pollution on agricultural productivity are not readily apparent until closely watched and analysed. Farmers often deal with challenges related to irritations and infections in their plants. Even though the aforementioned problem is being addressed, the detrimental impacts that are brought about by air pollution are not being addressed. The knowledge that is anticipated to differentiate and carry out these advancements is not practical nor available at a practicable level, even though the modifications in agricultural techniques may reduce the severity of these consequences[1].

Plant species that are distinct from one another react differently to pollution. Even though certain plant species may be able to withstand critical levels of pollution brought on by suspended particulate matter and buildup gases, other plant species are often susceptible to harm. As a result, how plants react to air pollution is contingent upon the many kinds of harmful compounds that are present, the quantity of those pollutants, and the extent to which they are susceptible to them [2].

[*] This work was supported by the Society for Industrial and Applied Mathematics

[†] Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India (shobanaj1@srmist.edu.in).

[‡] Department of Computer Science and Engineering, GITAM University, Bengaluru, India (vadhinar@gitam.edu).

[§] Department of Computer Science and Engineering, J N N Institute of Enginering, India and Saveetha university, Chennai, India (balamuruganp@jnn.edu.in).

[¶] Department of Computer Science and Engineering, Sasi Institute of Technology and Engineering, Tadepalligudem, Andhra Pradesh, India (sivakumarperumal@sasi.ac.in).

[‖] Department of Artificial Intelligence and Data Science K.Ramakrishnan College of Engineering, Samayapuram - Kariyamanickam Rd, Tamil Nadu, India. (sankarime09@gmail.com).

[**] PG and Research Department of Computer Science, Mary Matha Arts and Science College, Mananthavady, Kerala, India. (eldhokj@marymathacollege.ac.in).

[††] Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India. (nr7061@srmist.edu.in).

Urban areas are experiencing environmental pollution issues such water, noise, and air pollution due to economic and technological advancements. Air pollution directly affects human health by exposing individuals to toxins and particles, leading to a growing interest in its effect among scientists. The primary sources of air pollution are the combustion of fossil fuels, agricultural activities, emissions from factories and businesses, household heating, and natural calamities [2]. Air quality in the United States has been researched for the last thirty years with the establishment of the Clean Air Act programme. Despite the program's improvement in air quality over time, air pollution remains a persistent issue. About 200,000 premature deaths per year in the US are attributed to total combustion emissions, mostly caused by pollutants like particulate matter 2.5 (PM2.5), with an additional 10,000 fatalities per year linked to variations in ozone concentration.

The American Lung Association estimates that air pollution-related diseases cost about 37 dollars billion annually in the US, with California accounting for 15 billion dollars [7].Air pollution is the introduction of dangerous or excessive amounts of certain chemicals like gases, particles, and biological molecules into the atmosphere.

The high emissions result in detrimental effects such as infections, fatalities among people and other living species, and damage to crops. The primary air pollutants, referred to as criteria pollutants, include CO, SO2, lead, ground-level ozone (O3), NO2, and PM. The US Environmental Protection Agency (EPA) monitors the levels of these pollutants to regulate air quality. Scientific studies have shown a connection between brief exposure to these pollutants and various health issues, such as reduced capacity to meet higher oxygen requirements during physical activity (especially for individuals with heart conditions), airway inflammation in healthy individuals, heightened respiratory symptoms in asthma patients, respiratory crises in children and the elderly, and more.

The influence of air pollution on agricultural output is substantial, although frequently disregarded. While farmers tackle plant irritations and infections, the wider detrimental impacts of air pollution remain unattended to. Various plant species exhibit distinct responses to pollution, with certain species being more vulnerable to harm caused by pollutants such as particulate matter and gases. Urban areas are exposed to pollution from multiple sources, such as the burning of fossil fuels, emissions from industries, and agricultural practices. This pollution has a negative impact on human health and leads to premature deaths. Despite the implementation of initiatives such as the Clean Air Act, air pollution continues to be a significant problem. Machine learning (ML) provides potential solutions for assessing and reducing these effects by extracting valuable information from extensive datasets to forecast and enhance agricultural results.

Machine Learning(ML) is a subset of artificial intelligence(AL) that involves using statistical models to extract important insights from large datasets. The main difference between the method of statistical analysis and machine learning is that statistics focuses on representing data in numerical terms of probability or likelihood measures, rather than deterministic processes like cluster assignments, forecasting functions, etc. The assignments and tasks to be completed are roughly equivalent. The learning techniques are referred to as estimating strategies. Many researchers and analysts have already found that the basic concepts of machine learning closely resemble non-parametric estimation terminology [5]. ML allows systems to learn and improve from data without explicit programming. ML is distinct from AL and DL. ML's benefit is in training the model based on available data to predict future outcomes more efficiently. The subsections detail the individual characteristics of several supervised and unsupervised algorithms that impact agricultural frameworks.

The paper is structured as outlined below. Section 2 provides a thorough review of the literature, analysing past and relevant research. In Section 3, we present a suggested deep learning model, emphasising its predictive capabilities. Furthermore, provide a detailed description of the data used in this study and elaborate on the data pretreatment steps taken to create a more concise and informative dataset for analysis. Section 4 outlines our experimental investigation, including details of the experimental setup and analysis of the data. Section 5 closes the work and presents suggestions for further research.

2. Related Work. Saritha et al [3] the article underscores the harmful impact of air pollution on agriculture, noting that while the impacts may not be immediately apparent, they may greatly reduce agricultural productivity. The authors suggest using a machine learning method to examine the impact of air pollution on agriculture, focusing on the trends in crop production over time. This technique seeks to identify crop types that are more resistant by analysing pollution data, enabling farmers to make well-informed choices when selecting crops. The research explores the integration of air quality data with agricultural yield data to analyse the impact of contaminants on crop production. It indicates that certain crops have decreased production when exposed to high levels of pollutants, but others, such as sugarcane, are more resilient. The protocols offer an overview of available supervised and unsupervised machine learning models [4] connected with agricultural yield in literature. Highlighting the promise of machine learning in tackling difficult agricultural concerns including crop improvement, yield prediction, crop disease diagnosis, and recognizing water stress. Exploring the combination of agronomic elements with data analytic approaches to enhance crop yield forecasting. This might be helpful for agricultural academics and practitioners who want to use data-driven methods to enhance crop management and productivity.

The research [5] combines support vector regression (SVR) with a radial basis function (RBF) kernel to successfully estimate the concentrations of pollutants and the air quality index (AQI) in California. The authors show how employing the complete set of accessible variables is more successful than feature selection utilizing principal component analysis. The report offers future research areas, such as examining additional approaches for hyperparameter optimization and comparing SVR findings with additional algorithms for machine learning, which may lead to further breakthroughs in air quality prediction and simulation.

A complete survey of the current achievements in the application of deep learning (DL) in the agriculture industry. Highlights numerous uses of DL in agriculture, covering counting fruits, controlling water, crop management, soil management, weed identification, seed categorization, yield prediction, disease detection, and harvesting [6]. It underlines the problems encountered in applying DL in agriculture, including the complexity of assembling datasets, the expense of processing resources, and the scarcity of DL professionals. Addressing these difficulties may help overcome hurdles to the mainstream implementation of DL in agriculture.

The report provides an in-depth examination of recent developments and identifies areas for further investigation, such as robustness, interpretability, and integration of multiple data modalities. As such, it is an invaluable resource for future research and development in the field of deep learning in agriculture.

The study's methodology included a thorough examination of agricultural deep learning algorithms by analysing secondary data from academic publications released between 2016 and early 2022. Data collecting included using databases including Research Gate, IEEE Explore, Springer, Elsevier, Google Scholar, Frontier, and Science Direct. The focus was on scholarly journal articles and conference papers that were pertinent to the study goals. Studies predating 2016 were not included in the research. The paper utilised a range of deep learning tools for agricultural model development, such as Python tools for image saliency, gradient explanation technique, integrated gradient, DeepLIFT, guided backpropagation, class activation maps (CAMs), and layer-wise relevance propagation (LRP). The technologies were used to improve the precision and comprehensibility of deep learning models in the field of agriculture.

Assessing the current status of agricultural air quality research and pinpointing potential future research avenues to investigate contaminants associated with agriculture and their effects on air quality, human health, and regional climate. Developments in evaluations, modelling, emission controls, and farm operation management are necessary to successfully limit emissions from agriculture [7]. The significance of implementing laws and regulations to decrease agricultural emissions and their environmental effects. It is important to tackle the issues and uncertainties in present air quality models used in agriculture since doing so would enhance air quality, human health, agricultural settings, and biodiversity.

Examining recent studies and uses of artificial intelligence to lessen the negative impacts of climate change, particularly in fields like energy efficiency, carbon capture and storage, weather and renewable energy prediction, grid control, architectural design, transportation, precision farming, industrial operations, deforestation reduction, and sustainable urban development.AI can play a crucial role in mitigating the effects of climate change by improving energy efficiency, decreasing energy usage in buildings, and optimizing power systems to lower electricity costs [8]. Integrating AI with smart grids can enhance the efficiency of power systems, resulting in less energy wastage and reduced electricity costs. AI integrated with transportation systems can decrease carbon dioxide emissions by around 60 percent. AI can assist in the conservation of natural resources by decreasing deforestation and encouraging sustainability. It can also help in designing resilient cities to reduce damage from severe weather events.

In all history, humans have depended on intuition, shared knowledge, and sensory cues to make successful

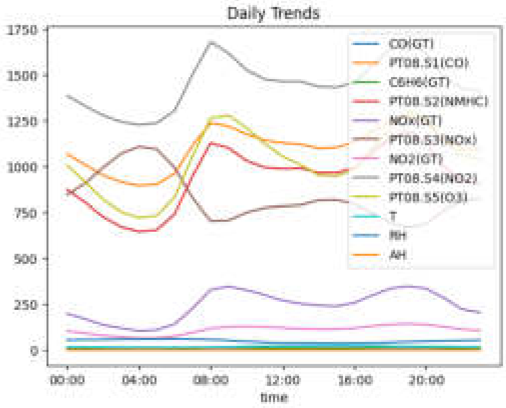


Fig. 2.1: Daily trend

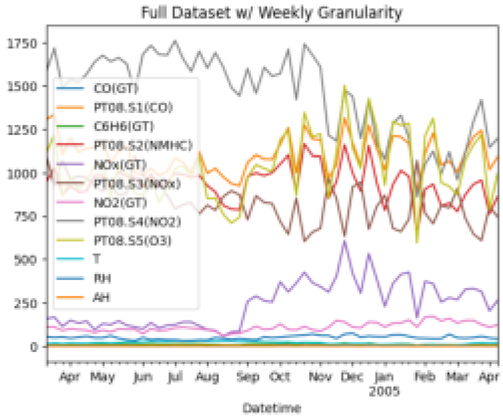


Fig. 2.2: Weekly prediction

decisions in animal husbandry since the early days of domestication. This has significantly improved our practices in animal husbandry and agriculture. The increasing need for food and the progress in sensing technologies could enhance the centralization, size, and efficiency of animal farming. It can revolutionize animal agriculture. This study delves into the challenges and opportunities posed by sensor technology in assisting animal producers to increase meat and animal product production. This study [9]delves at how sensors, big data, artificial intelligence, and machine learning may assist animal producers in reducing production costs, improving efficiency, enhancing animal welfare, and increasing the number of animals per hectare. The text delves into the difficulties and constraints of technology.

On the figure 2.1, the x-axis reflects the passage of time, while the y-axis depicts the level of the substances expressed in parts per billion (ppb). The concentrations of the majority of compounds (CO(GT), C6H6(GT), NOX(GT), and NO2(GT)) tend to change during the course of the day, with values that are typically greater in both the morning and the evening and values that are lower in the afternoon. PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), and PT08.S4(NO2) all exhibit comparable patterns, which suggests that they may be co-located sensors that measure the same air quality conditions. Not only can temperature (T) and relative humidity (RH) fluctuate during the day, but they also have the capacity to affect the quantities of certain compounds.

Figure 2.2 illustrate the timeseries graph, it shows the the daily fluctuations of various air quality measures over the year. These data were most likely gathered at a monitoring station. The date is displayed along the

Table 2.1: Data Sources and Collected Posts

Author	**Methods**	**Contribution**	**Limitations**
Saritha et al [3]	Machine learning	Examines the impact of air pollution on agriculture by analyzing trends in crop production over time	Limited to data availability and the complexity of modelling the relationship between air pollution and crop yields
Elavarasan et al [5]	Support vector regression with radial basis function kernel	Estimates the concentrations of pollutants and the air quality index in California	Limited to the specific case study of California and may not be generalizable to other regions
Albahar et al [6]	Deep learning	Identifies various applications of deep learning in agriculture, including fruit counting, water control, crop management, and disease detection	Highlights challenges such as data complexity, processing resource requirements, and scarcity of deep learning expertise
Aneja et al [7]	Air quality modeling and assessment	Assesses the current status of agricultural air quality research and identifies potential future research avenues	Limited by the uncertainties and complexities involved in air quality modeling
Chen et al [8]	Artificial intelligence (AI)	Examines the use of AI for climate change mitigation and adaptation, including applications in energy efficiency, renewable energy, and sustainable agriculture	Limited by the need for further research and development to address specific challenges and ensure ethical implementation

x-axis, which extends from April of the previous year to March of the current year. A number of different contaminants are represented along the y-axis, and their concentrations are expressed in parts per billion (ppb). As the year progresses, it appears that the concentrations of the majority of pollutants (CO(GT), C6H6(GT), NOX(GT), and NO2(GT)) change. It is possible that the values are greater during the months of October through March as compared to the months of April through September when temperatures are higher. It's possible that this pattern is the result of seasonal shifts in the weather conditions or actions carried out by humans that have an effect on these pollutants. Readings from ground-truth instruments (CO(GT), NOX(GT), and NO2(GT)) and perhaps related sensor readings from particular places (PT08.S1(CO), PT08.S3(NOx), and PT08.S4(NO2)) are included in the measurements for some pollutants, such as carbon monoxide, nitrogen oxides, and nitrogen dioxide. It would appear that the patterns between these values are comparable, which would imply that the sensors are catching circumstances of air quality that are equivalent to one another. In addition, graphs for temperature (T) and relative humidity (RH) are now included in the graph. It is possible for certain environmental conditions to have an effect on the concentrations of some contaminants in the atmosphere.

3. Research Methods.

3.1. Prediction model. Air pollution is the presence of polluting substances that contaminate the air. Air pollution often involves the presence of solid, liquid, and gaseous particles in the outside world. These particles are emitted by fuel and petroleum in automobiles, waste produced from companies or industries in liquid or gas form, ashes produced by volcanoes or wildfires, burning of garbage and fossil fuels, and other sources. Air pollution in real-time is a significant factor in causing chronic illnesses in humans and impacting the environment's natural resources [10]. It also affects the agriculture sector by hindering appropriate crop growth and reducing the productivity of farming. Long-term health issues include nerve damage, lung cancer from inhaling polluted air, kidney failure, and many child health problems. On a global scale, it leads to consequences including ozone layer depletion, acid rain, and global warming, which in turn cause reduced rainfall. A novel ensemble learning model based on a meta-heuristic algorithm is proposed to address the identified issues and improve air quality prediction outcomes.

Generally, the involves incorporating several deep-learning techniques to improve performance. This research utilizes learning techniques such as Bi-LSTM to construct the ensemble model. The methods are executed similarly to a neural network, with neurons capable of categorizing the factors that lead to the outcomes in AQ.

Automated systems are highly proficient at swiftly gathering, handling, and evaluating huge amounts of data. They are unable to make efficient decisions in the absence of data. They can help humans improve decision-making by gathering and analyzing vast amounts of comprehensive data. Various sensors can assist farmers in monitoring animal activities in real-time on a farm. Sophisticated algorithms [11]can utilize large datasets to monitor, measure, and comprehend alterations in animal behavior. Consequently, this can assist farmers in making more informed decisions and implementing timely disease interventions. Air sensors in the poultry sector can now anticipate the beginning of Coccidiosis, an intestinal infection that may rapidly spread among birds without showing any visible signs. One method to detect this illness is by consistently observing air quality. The concentration of volatile organic compounds (VOC) in the air rises with the increasing number of diseased birds. Air sensors can notice this shift earlier than a farmer or doctor. Once the farmers are informed, they can promptly implement measures to halt the illness from spreading. This technique conserves multiple animal lives and averts financial damages.

3.2. Preprocessing. Preprocessing is commonly employed to eliminate redundant features in order to enhance performance. The air data Az is inputted into the first step of the model suggested as a preprocessing strategy. The supplied data typically contain missing values, outliers, and redundant data. Preprocessing is utilized to address these limitations in order to improve the accuracy of the model. Data preparation involves data imputation, data cleansing, and data transformation [12].

Data imputation is utilized to address the absence of values in the input data Az. Missing data are either replaced with a value of zero or estimated using the mean value of the entire sample and the nearest available data point. The data is imputed using an arbitrary data sample represented as A_z^{imp}.

This strategy is employed to identify and eliminate errors and discrepancies. Invalid input. The input data comprises noisy data, outliers, undesirable qualities, and irrelevant data. High computational time occurs when working with irrelevant data, noise, or outliers that lead to errors and inconsistent analysis [13]. Data cleansing is utilized to eliminate redundant data in order to address these issues, resulting in improved performance accuracy and reduced computation time. The resulting data are as follows A_z^{cle}.

Data transformation involves standardizing and consolidating the data. Transforming information is utilized to convert one format of environment data, which includes various sorts of particles such as solid, liquid, and gas, into another one. [14]Transforming the data facilitates predicting air quality and improving performance analysis. The result is what it comes from. The ultimate preprocessed data $A_z^{\mathrm{tra}\,a}$ is then sent on to the feature extraction stage.

3.3. Deep Learning - CNN. The first application of convolutional neural networks is in picture data processing. A multiple-layer perceptron network containing many hidden layers is the structure of a deep network of convolutional neural networks [15][16]. The layer of convolution consists of artificial neurons that represent convolutional filtering and are used to construct feature maps [17]. It is necessary to divide the input into smaller blocks in order to convolve it using a particular set of weights. Through the application of convolutional filters with the same weights to the input, several sets of features may be produced. The pooling layer is applied in order to reduce the number of parameters as well as the dimension of space of the provided data representation. This is accomplished by minimizing the number of parameters. It has been established which is data that is similar in the particular region, because the reaction that is most prevalent is output. An example of a nonlinear function that is utilized with the purpose of learning complex nonlinear structures is the activation functionThe study's learning layers make use of both the exponential and the Rectified Linear Unit (ReLU) [18]. Feature aggregation through global examination of outputs from preceding layers is accomplished by this fully linked learning layer, which is located at the very end of the neural network.

3.4. Long short-term memory. It is the sequence of inputs that determines the output of neural networks with recurrent neurons (RNNs), and the network creates different outputs according to the same input regardless of the order in which the inputs are presented. RNNs combine information from the past with the information that is now being processed during the generation of the output. Long short-term memory, also known as LSTM, is a specialized kind of recurrent neural networks (RNNs) that is employed for the purpose of identifying the long-term dependencies present in sequence data. Data is received externally, stored, recorded in memory cells, and accessed through gates. The memory unit is responsible for controlling the flow of information in order to determine the impact that prior information has on output. Additionally, this unit stores a copy of the predictions that have been made.

After multiplying the weights and information stored in memory, a decision is made on which data will be utilized and how much of it will be used. Some of the weights and information are subsequently added again to the forecast. Some forecasts are chosen as the current prediction, while irrelevant information is isolated to prevent it from influencing future predictions.

The equations for a Long Short-Term Memory (LSTM) cell are as follows:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{3.1}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{3.2}$$

$$\tilde{C_t} = \tanh(W_{ic}x_t + b_{ic} + W_{hc}h_{t-1} + b_{hc}) \tag{3.3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C_t} \tag{3.4}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{3.5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{3.6}$$

Here, x_t is the input at time t, h_t is the hidden state at time t, C_t is the cell state at time t, W and b are weight matrices and bias vectors, σ is the sigmoid activation function, tanh is the hyperbolic tangent activation function, and $\odot$ represents element-wise multiplication.

3.5. Implementation. The initial step of data preparation involved cleaning the data and, in the event that any values were missing for up to five successive time periods, interpolating the data.The minimum-maximum normalization approach was then used to standardize the data.

A set of interpolated time series measurements was constructed using a variety of frame sizes and a number of different data separation techniques.Following the definition of both two-dimensional and three-dimensional input structures,Through the utilization of the Deep Neural Designer Tool that is incorporated into MATLAB edition R2020a, a CNN+LSTM deep learning–based period forecasting framework was constructed. A number of hyper-parameters were adjusted in order to improve the predictive capability to make accurate predictions. Throughout both training and testing, the hyper-parameters were tuned to optimize their performance. It was determined that the structure of the neural network that had the least validating RMSE was the one that should be utilized on test data after the algorithm was executed fifteen times for each approach. Alterations were made to the characteristics and hyperparameters of the neural network, including the hidden layer's type and quantity, the total number of neurons, and the activation function. This allowed for the neural network to be rebuilt and trained.

The Root Mean Square Error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.7}$$

In the following equation:

n is the number of data points or observations.

$\hat{y}_i$ is the predicted value for the i-th observation.

y_i is the true or observed value for the i-th observation.

$\sum_{i=1}^{n}$ denotes the summation over all data points.

Table 3.1 illustrates the comparision of the different model with the evaluation metrices.

Table 3.1: Root Mean Square Error (RMSE) for Different Pollutants

Model	RMSE
LSTM	0.47
BiLSTM	0.25
Proposed Ensemble LSTM+CNN	0.16

3.6. Hyperparameter tuning. Hyperparameters may be associated with the tasks of model selection, such as the topology and size of the network [19], or the pace of learning as well as the size of the mini-batch are two examples of the factors that may be related with the technique for optimizing and the learning procedure among these factors.

This study focused on tuning a specific set of hyperparameters:

Frame Size: The frame size was adjusted using integer values within a specific range.

Step Size: In this investigation, a step size of 1 was used for each time period sample.

Splitting the Data: The training set, validation set, and test set have their data separated by 70%-15%-15% and 80%-10%-10%, respectively.

Selection of Samples: There are three ways to choose a sample: randomly, sequentially, or consecutively. When data samples for training, validation, and testing are chosen at random, this is known as the random selection approach. When data is divided into training and test sets in order, sequential sample selection is used to determine with the use of an evaluation rate, verification samples generated from the original data set. As part of the sequential selection process, the data is broken up into three distinct sections: training, verification, and test procedures.

Validation Frequency: Validation rate is the amount of iterations that occur between assessments of validation metrics. The three values that were taken into consideration for validation frequency were 10, 15, and 20, accordingly. Additionally, it is a reference to the time period during which the validation sample is selected using the ordered choice of sample method.

Sizes of mini-batch: The number of iterations that were selected to be 20, 80, 110, 160, and 200 was done so with the intention of achieving learning progress. This was determined based on the volume of the time series data collection.

Number of epoch: When determining when to stop training the model, this study made use of the early stopping approach.

The number of layers that are convolutional: There were three levels of convolution used in the construction of the CNN part: one, two, and three layers collectively.

Layers of Pooling: When the CNN module was initially developed, there was no pooling layer included in its building process. Following that, a maximum pooling layer and an average pooling layer were applied in the process of building the structure.

Activations Functions: In this particular examination, the Rectified Linear Unit (ReLU) with sigmoid equation were both applied as activation functions simultaneously.

Figure 3.1 illustrated the correlation matrix from an air pollution experiment shows the connections among different pollutants. The correlation coefficient in each cell represents the degree to which two metrics fluctuate together: an amount near 1 indicates a high tendency to rise or decrease in a similar manner, while a value around 0 indicates a weak or missing link. Analyzing these correlations can assist researchers in comprehending the relationships between contaminants and their possible sources or methods of reduction.

CO(GT): Carbon Monoxide (Ground Truth) - perhaps a standard measurement for Carbon Monoxide.
PT08.S1(CO): Probably a Carbon Monoxide sensor reading from a particular position (PT08.S1)
Non-Methane Hydrocarbons (Ground Truth): NMHC(GT)
C6H6(GT): Benzene (Ground Truth)
NOX(GT): Nitrogen Oxides (Ground Truth)
PT08.S2(NMHC): Sensor reading indicating Non-Methane Hydrocarbons

Fig 3.2 shows the a correlation matrix from a study involving partial NMHC (Non-Methane Hydrocarbon)

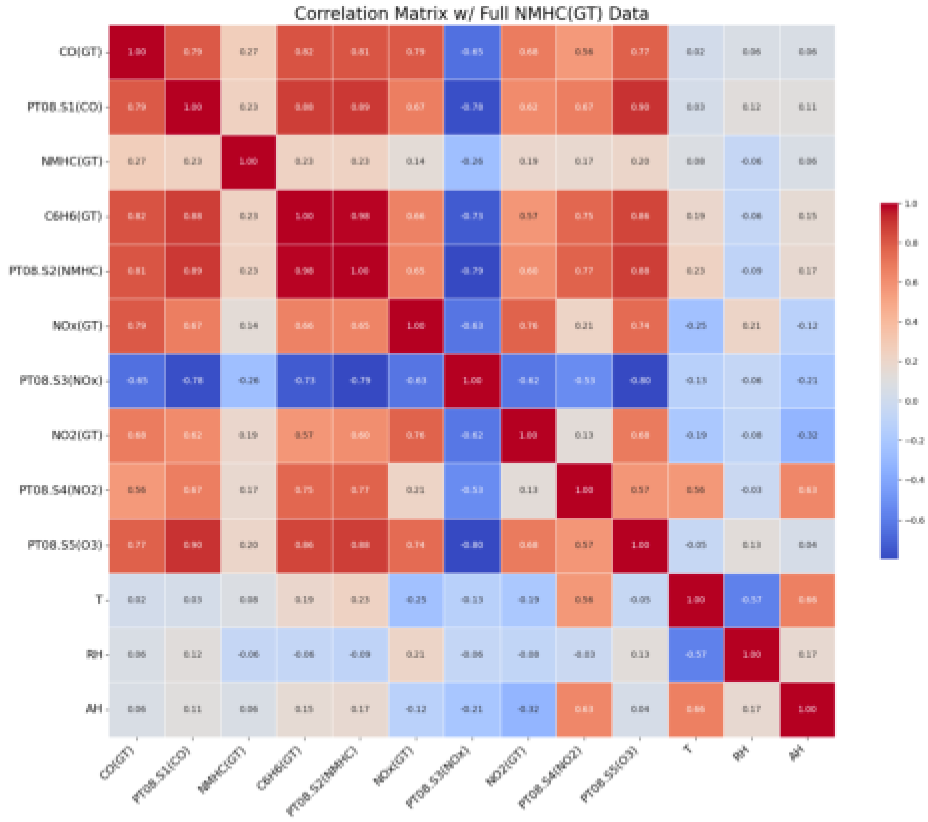


Fig. 3.1: Correlation Matrix

data. The "partial NMHC(GT) data" indicates that the data for non-methane hydrocarbons are inadequate. This might be the result of sensor readings that are either missing or incorrect. Through an analysis of these connections, analysts are able to get an understanding of the ways in which pollutants and other factors may be connected to one another and impact one another.

The heatmap is an illustration in the figure 3.3. The correlation coefficients that exist between air quality indicators and any additional variables that may be significant. A pair of variables, such as carbon monoxide and nitrogen oxides, temperature and humidity, is represented by each individual cell. The degree of correlation and the direction of the correlation are both indicated by the intensity of the color in the cell.

4. Results and Discussions. In pooling layer, which reduces the amount of information that is included in the enormous quantity of input data, the key features of the input are lost, and the size of the input is reduced. In order to make an accurate prediction of the quantity of contaminants using each approach, the model was executed fifteen times for each and every possible arrangement of the hyper parameter [20] variables specified previously. Finally, use the equation 3.7 and the test RMSE as well as the correlation values were computed after the most effective test results were selected based on the values of the RMSE that were the lowest.

During the training phase, the RMSE was found to be at its lowest when the learning rate was set at 0.005. Additionally, the random sample selection approach achieved a lower random sample standard error (RMSE) significance in comparison to the sequential and consecutive techniques of sample selection. As the metric for evaluating the performance of the prediction, the test RMSE values were utilized, and the best possible prediction performance was attained.

Figure 4.1 shoes the scatter plot and show the expected values compared to the actual values of errors,

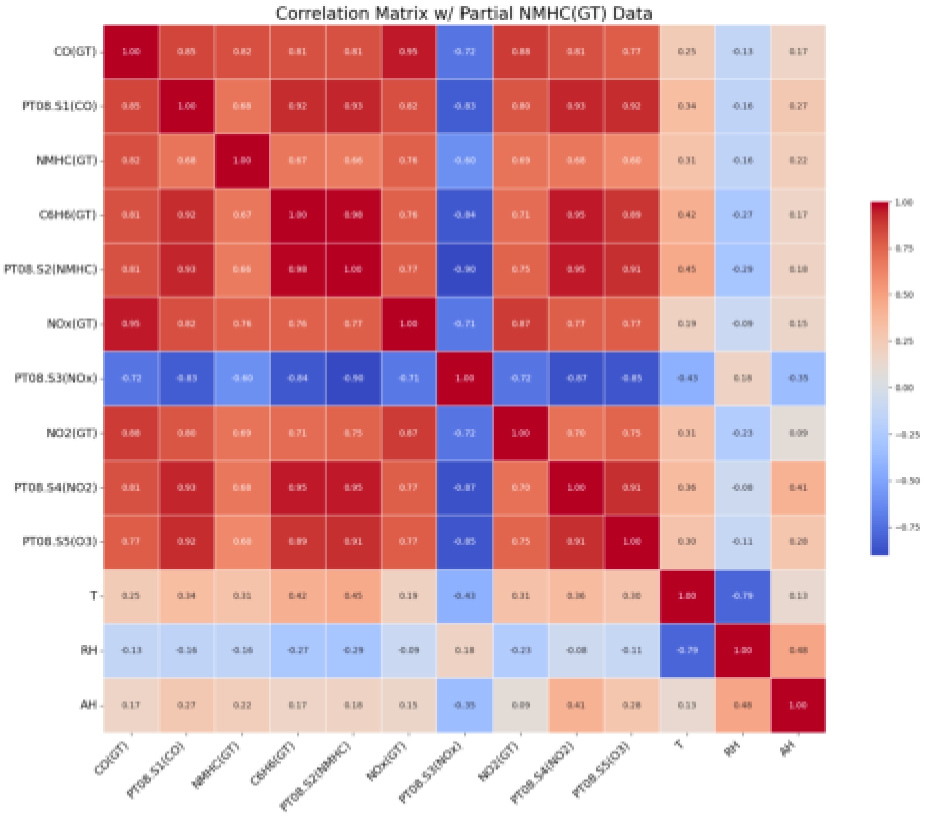


Fig. 3.2: Correlation Matrix

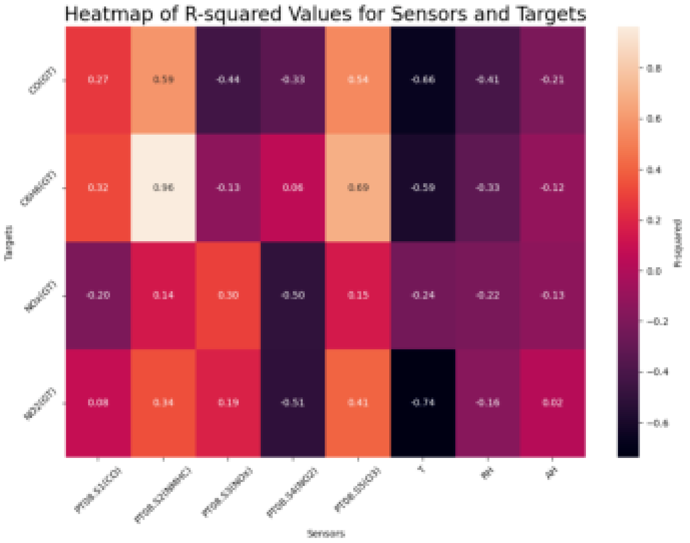


Fig. 3.3: Heat Map

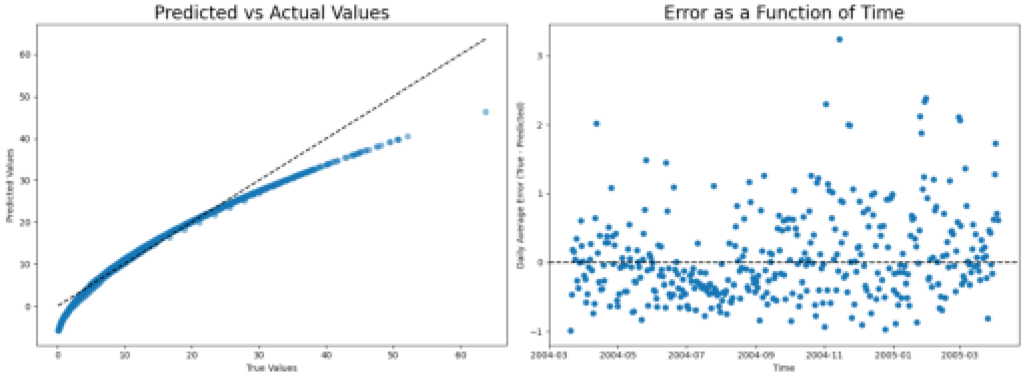


Fig. 4.1: Prediction

as well as the error plotted against time, likely from an air quality study. The x-axis depicts a timeline from March 2004 to March 2005. The left y-axis depicts the error, while the right y-axis reflects the expected and actual values. The graph indicates that the forecasted values consistently exceed the real values, especially in the initial portion of the time frame. The error over time indicates that the forecasted values consistently exceed the real values, with the error diminishing as time progresses.

5. Conclusion. The rapid growth of the industry has led to a concerning problem of air pollution negatively impacting agricultural regions. We have chosen to create an application that predicts the best crop to minimise negative consequences based on existing pollution data. Therefore, the proposed hybrid Deep learning model to predict the air pollutants, and the evaluation metrices Root Mean Square error gives the low error value is 0.16. When constructing the Long Short term Memory model combination with the convolutional neural network in deep learning, we have taken into account the available contaminants. By entering geologically related data, the programme can properly analyse and anticipate the outcome.

REFERENCES

[1] Al-Janabi S, Mohammad M, Al-Sultan A. *A new method for prediction of air pollution based on intelligent computation,Soft Computing [Internet]. 2019 Nov 28;24(1):661–80, 10.1007/s00500-019-04495-1*

[2] Li Y, Sha Z, Tang A, Goulding K, Liu X *The application of machine learning to air pollution research: A bibliometric analysis. Ecotoxicology and Environmental Safety, 2023 Jun;257:114911, dx.doi.org/10.1016/j.ecoenv.2023.114911*

[3] Saritha and Shetty, Ritika and Devi, Manasa and Dhotre, Akash and Hanchinal, Prema R, *Analysis on Air Quality and its Effects on Agriculture*, 2022 3rd International Conference for Emerging Technology (INCET), 2022.

[4] Castelli, Mauro and Clemente, Fabiana Martins and Popovič, Aleš and Silva, Sara and Vanneschi, Leonardo, *A Machine Learning Approach to Predict Air Quality in , Complexity, Hindawi Limited, 2020.*

[5] Elavarasan, Dhivya and Vincent, Durai Raj and Sharma, Vishal and Zomaya, Albert Y. and Srinivasan, Kathiravan, *Forecasting yield by integrating agrarian factors and machine learning models: A survey*

[6] Albahar, Marwan, *A Survey on Deep Learning and Its Impact on Agriculture: Challenges and Opportunities, 13, 2077-0472, Agriculture, MDPI AG, 2023 10.3390/agriculture13030540*

[7] Aneja, Viney P. and Schlesinger, William H. and Erisman, Jan Willem *Effects of Agriculture upon the Air Quality and Climate: Research, Policy, and Regulations Environmental Science Technology, vol. 43, no. 12, pp. 4234–4240, May 2009, doi: 10.1021/es8024403.*

[8] Chen, Lin and Chen, Zhonghao and Zhang, Yubing and Liu, Yunfei and Osman, Ahmed I. and Farghali, Mohamed and Hua, Jianmin and Al-Fatesh, Ahmed and Ihara, Ikko and Rooney, David W. and Yap, Pow-Seng *Artificial intelligence-based solutions for climate change: a review, 2021, 10.1007/s10311-023-01617-y*

[9] *“.Deep Learning and Machine Learning based Air Pollution Prediction Model for Smart Environment Design Planning. Global NEST Journal , 2023” n.d. http://dx.doi.org/10.30955/gnj.004735.*

[10] Devasekhar V, Natarajan P. *Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques. International Journal of Advanced Computer Science and Applications, 2023;14(4), http://dx.doi.org/10.14569/ijacsa.2023.01404103*

[11] Méndez M, Merayo MG, Núñez M. *Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review. 2023 Feb 16;56(9):10031–66. http://dx.doi.org/10.1007/s10462-023-10424-4*

[12] Zadmirzaei M, Hasanzadeh F, Susaeta A, Gutiérrez E. *A novel integrated fuzzy DEA–artificial intelligence approach for assessing environmental efficiency and predicting CO2 emissions. Soft Computing. 2023 May 12;28(1):565–91. http://dx.doi.org/10.1007/s00500-023-08300-y*

[13] Sonawani S, Patil K. *Air quality measurement, prediction and warning using transfer learning based IOT system for ambient assisted living. International Journal of Pervasive Computing and Communications. 2023. http://dx.doi.org/10.1007/s00500-023-08300-y*

[14] Singh RB, Patra KC, Pradhan B, Samantra A. *HDTO-DeepAR: A novel hybrid approach to forecast surface water quality indicators. Journal of Environmental Management. 2024. http://dx.doi.org/10.1016/j.jenvman.2024.120091*

[15] Anyaene IH, Onukwuli OD, Babayemi AK, Obiora-Okafo IA, Ezeh EM. *Application of Bio Coagulation–Flocculation and Soft Computing Aids for the Removal of Organic Pollutants in Aquaculture Effluent Discharge. Chemistry Africa. 2023. http://dx.doi.org/10.1007/s42250-023-00754-9.*

[16] Foreback B, Mahura A, Clusius P, Xavier C, Baykara M, Zhou P, et al. *A new implementation of FLEXPART with Enviro-HIRLAM meteorological input, and a case study during a heavy air pollution event. Big Earth Data. http://dx.doi.org/10.1080/20964471.2024.2316320.*

[17] Lakshmipathy M, Prasad MJS, Kodandaramaiah GN. *Advanced ambient air quality prediction through weighted feature selection and improved reptile search ensemble learning. Knowledge and Information Systems. 2023. http://dx.doi.org/10.1007/s10115-023-01947-x*

[18] Janarthanan R, Partheeban P, Somasundaram K, Navin Elamparithi P. *A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society. 2021. http://dx.doi.org/10.1016/j.scs.2021.102720*

[19] Zhang Z, Johansson C, Engardt M, Stafoggia M, Ma X. *Improving 3-day deterministic air pollution forecasts using machine learning algorithms. Atmospheric Chemistry and Physics. 2024. http://dx.doi.org/10.5194/acp-24-807-2024.*

[20] Gholami H, Mohammadifar A, Behrooz RD, Kaskaoutis DG, Li Y, Song Y. *Intrinsic and extrinsic techniques for quantification uncertainty of an interpretable GRU deep learning model used to predict atmospheric total suspended particulates (TSP) in Zabol, Iran during the dusty period of 120-days wind. Environmental Pollution. 2024 http://dx.doi.org/10.1016/j.envpol.2023.123082.*

Edited by: Vadivel Ayyasamy

Special issue on: Internet of Things and Autonomous Unmanned Aerial Vehicle Technologies
for Smart Agriculture Research and Practice

Received: Jun 26, 2024

Accepted: Sep 12, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1017–1034, DOI 10.12694/scpe.v26i3.4166

RESEARCH ON IMPROVED RBM RECOMMENDATION ALGORITHM BASED ON GIBBS SAMPLING

QIAN XIAODONG[*] AND LAN JIABAO[†]

Abstract. Restricted Boltzmann Machine (RBM) is an important tool for personalized recommendation prediction, but it ignores the power-law distribution of the Restricted Boltzmann Machine data set, the RBM algorithm can not focus on the tail data sampling of the recommended data set. Therefore, firstly, the recommended data are obtained and the data characteristics are analyzed, then the random Gibbs Sampling initial value of RBM is changed to random selection in the early iteration and the last sampling value in the later iteration, the fixed Gibbs sampling steps were replaced by single-step sampling (CD-1) and multi-step sampling (CD-5),which is Periodic Gibbs Sampling (PGS). The experiment shows that the improved Gibbs sampling initial value and the changed Gibbs sampling steps can effectively improve the sampling performance, the improved RBM algorithm is also more accurate than the original RBM algorithm, the cyclic time Restricted Boltzmann Machine (RTRBM) algorithm and the Probability Matrix Factorization (PMF) algorithm. It shows that the improved RBM algorithm is suitable for the power-law distribution of recommendation data sets, and effectively improves the accuracy of recommendation.

Key words: Recommendation Algorithm, RBM model, Gibbs Sampling, power-law distribution polynomial

1. Introduction.

With the rapid development of e-commerce, the scale of consumer behavior data has been growing exponentially, making it difficult for consumers to locate satisfactory products among the vast amounts of product data. Recommendation algorithms, known for their simplicity and robustness, have become indispensable tools for assisting in recommendation decisions. To address the problem of ”information overload”, many companies use recommendation algorithms to intelligently mine and predict large-scale recommendation data, thereby increasing user engagement and consumption. However, as the complexity of recommendation data increases, it becomes challenging to balance the efficiency and accuracy of recommendation algorithms. Therefore, improvements to the efficiency and accuracy of the algorithms themselves are necessary.

Regarding the necessity for improvements in recommendation algorithms, this study focuses on the Restricted Boltzmann Machine (RBM) recommendation algorithm, based on the theories of complex networks and Markov chains. Considering the characteristics of recommendation data, the research aims to summarize and improve the parameter iteration algorithm of RBM, specifically the Gibbs sampling principle, to explore a more reasonable and efficient RBM algorithm. Comparative experiments between the improved RBM recommendation algorithm and the original RBM will be conducted to provide feasible references for the enhancement of recommendation algorithms.

2. Literature Review.

As a widely used neural network model in practical applications, the RBM algorithm is capable of making effective recommendation predictions in recommendation scenarios. Therefore, it has been extensively studied by scholars both domestically and internationally. The analysis of improvements to the RBM recommendation algorithm can be primarily divided into two aspects: recommendation data and recommendation algorithms. The improvements are further categorized into the characteristics of recommendation data, the integration and enhancement of the RBM algorithm, and the inherent improvements of the RBM algorithm itself. Each of these aspects will be introduced and evaluated in detail.

2.1. Analysis of Recommendation Data Characteristics.

Before conducting research on recommendation algorithms, analyzing the characteristics of recommendation data or the networks they form, and quantifying these features within the recommendation algorithms, can improve the efficiency of algorithm enhancements. Consequently, many scholars have carried out relevant research in this area. Garima and Rahul[1]

[*] School of Economics and Management, Lanzhou Jiaotong University, Lanzhou Gansu 730070, China (qianxd@mail.lzjtu.cn).
[†] School of Transportation, Lanzhou Jiaotong University, Lanzhou Gansu 730070, China

used text mining and sentiment analysis to extract relevant information from the text information of food and aperitif wine, and concluded the power law characteristics of the data set.Ralph and Patrycja[2] analyse the characteristics of the data set based on two-dimensional image, the relevant scaling parameters are extracted accurately, and the power-law distribution is proved. Fang et al [3] proposed a contrastive meta-learning framework (CM-HIN) based on heterogeneous information networks. This framework utilizes meta-paths and network motifs to capture both high-order and local structure information of heterogeneous information networks, thereby improving the precision of recommendation network construction. Wang et al [4] also noted the heterogeneity in recommendation data and proposed a commodity recommendation framework based on self-attention mechanism for attribute heterogeneous information network embedding. This framework learns the latent information contained in different edge types and attribute embeddings to increase the effective information in the recommendation network.

Besides focusing on the heterogeneity of recommendation networks, many scholars also delve into the latent features of these networks. For instance, Ambikesh et al [5] proposed methodology, termed GOA-k-means, amalgamates the Grasshopper Optimization Algorithm (GOA) with k-means clustering to navigate the dynamic nature of user preferences. Facilitating real-time calibration, GOA-k-means yields recommendations that adapt to users' shifting interests. By combining neural network Doc2vec and word bag BOW, Hafez et al [6] construct a multi-standard recommendation system.

2.2. Integration and Improvement of the RBM Algorithm. In the process of recommendation prediction, most studies on recommendation systems do not distinctly categorize improvements into recommendation data and recommendation algorithms. For instance, Jha et al [7] proposed a hyper-tuned Restricted Boltzmann Machine (RBM), using a contrastive divergence learning algorithm to regenerate tabular data models for enhancing recommendation accuracy. Harshvardhan et al [8] introduced a time-aware recommendation system based on unsupervised Boltzmann Machines (UBMTR) to detect latent hidden features related to the time of each rating in user movie rating data. Fachechi et al [9] calculated the relevance of recommendation data and constructed intra-layer connections for the neurons in the hidden layer of the RBM, thereby creating the Dream Boltzmann Machine (DBM). Xie et al [10] extracted user and resource features from the recommendation system to construct a multi-layer RBM network, forming a deeply stable personalized recommendation model using the Restricted Boltzmann Machine to compute recommendation results. Wu et al [11] proposed an improved hybrid recommendation algorithm based on Gaussian RBM. They used a convolutional neural network to obtain latent feature vectors of text information and rating information, merged user vectors and item vectors into a user-item matrix, and input this matrix into the visual layer of the Gaussian RBM to predict ratings.

From the aforementioned literature, it is evident that most personalized recommendation approaches using the RBM model have focused on improving the recommendation datasets or combining RBM with other algorithms, without addressing the intrinsic time costs, accuracy, and other aspects of the RBM model itself. Therefore, further research on the RBM recommendation algorithm needs to supplement and refine these aspects to enhance its overall performance.

2.3. Improvements to the RBM Algorithm. To enhance the performance of the RBM algorithm, scholars have started focusing on improving its efficiency and accuracy, particularly in two main areas: the initialization of sampling values and the optimization of parameter gradients. The original Gibbs sampling initialization uses the initial training sample values, but randomly selected training samples can lead to increased training time costs and reduced accuracy. Therefore, Tieleman [12], building on the CD algorithm, proposed the Persistent Contrastive Divergence (PCD) algorithm. PCD used the sampled values from the previous iteration as the initial values for the next sampling iteration, accelerating the convergence speed of the Gibbs sampling chain. To further speed up the PCD algorithm, Tieleman [13] introduced the Fast Persistent Contrastive Divergence (FPCD) algorithm, which includes additional acceleration parameters to enhance sampling speed.

To improve the effectiveness of initial value selection, Li et al [14] proposed the Dynamic Initial Value Algorithm (DIS), which dynamically improves the initial values for Gibbs sampling. Savitha et al [15] introduced the Online Restricted Boltzmann Machine (O-RBM), which adjusted the initial values for Gibbs sampling, constructing a probability distribution of data information to achieve unsupervised learning for recommendation predictions.

In the realm of optimizing parameter gradients in Gibbs sampling, Li et al [16] conducted an analysis of the numerical and directional errors between the approximate gradient and the true gradient of the RBM model. They devised two algorithms to mitigate these errors: the Gradient Fixed Gibbs Sampling (GFGS) training algorithm and the Gradient Fixed Parallel Tempering (GFPT) algorithm. These methods aim to adjust the numerical values and directional aspects of the approximate gradient, thereby reducing errors during training. Ma and Wang [17] identified biases in the parameter Gibbs sampling of RBM models, which fail to achieve maximum likelihood parameters. Leveraging the principle that the average of random variables approximates the expected value, they introduced the Averaged Contrastive Divergence (ACD) algorithm to mitigate the bias in maximum likelihood parameters. Kirubahari and Amali [18] utilized Bayesian Optimization (BO) to enhance the hyperparameters of Restricted Boltzmann Machines. By optimizing the number of sampling steps, they aimed to improve prediction quality. Wang et al [19] proposed the Three-Phase Gibbs Sampling (PGS) method, which involves training RBMs using different data distributions across phases to achieve more effective parameter extraction and feature reconstruction.

In addition to these advancements, research on the number of Gibbs sampling steps in RBM algorithms remains relatively limited. Li et al [14] conducted detailed research on the selection of sampling steps. However, since the choice of Gibbs sampling steps significantly impacts the time cost and training accuracy of RBM models, it remains a crucial area requiring further investigation.

2.4. Literature Review Summary. Through the analysis of the aforementioned studies, it is evident that many scholars analyze recommendation data and then quantify these features into recommendation algorithms to improve recommendation accuracy. However, due to the vastness of recommendation data and the inherent accuracy limitations of recommendation algorithms, current research still has several shortcomings. Firstly, most studies on recommendation data construct data networks and then investigate the characteristics of these networks based on their properties. However, when the characteristics of recommendation data are difficult to quantify within a network, it becomes challenging to construct a network that accurately reflects these characteristics to obtain meaningful insights from the recommendation data [1, 2, 3, 4, 5, 6]. And then, although the RBM model is widely used in the field of recommendations, most research [7, 8, 9, 10, 11]focuses on improving the data inputted into the RBM model rather than enhancing the operational speed or accuracy of the RBM model itself. Even within studies aimed at improving the RBM model [16, 17, 18, 19], which tends to overlook issues such as the problem of important data information not being learned due to the random initialization of Gibbs sampling, as well as the drawback of fixed sampling steps, which makes it difficult to improve prediction accuracy in the later stages of algorithm iteration.

To address the first issue, user social attention information is statistically analyzed to obtain the power-law distribution characteristics of the recommendation data. To tackle the second issue, the Gibbs sampling approach is adjusted by incorporating these power-law distribution characteristics. Specifically, the initial values of Gibbs sampling are set to be random in the early stages of iteration and are replaced by the previous sampling results in the later stages. Additionally, fixed Gibbs sampling steps are replaced with periodic Gibbs sampling (PGS).

3. Analysis of RBM Algorithm Improvements. Most recommendation datasets exhibit a power-law distribution, indicating that the recommendation data is primarily concentrated in the tail [20, 21]. The main algorithm in RBM (Restricted Boltzmann Machine) is Gibbs sampling, where the initial values are randomly selected from the recommendation data. This random selection fails to focus on tail data, lacking deep iterative analysis of tail data and not aligning with the long-tail characteristics of recommendation networks. Similarly, Gibbs' fixed number of sampling steps processes both the head and tail of the recommendation dataset with the same number of steps, collecting an equal amount of recommendation data. This approach does not allow for concentrated learning of tail information, resulting in insufficient learning and representation of tail user information.

Therefore, the Gibbs sampling method will be improved in terms of sampling initial values and sampling steps to enhance the recommendation performance of the RBM algorithm. The process will be as follows:

1. Provide a brief overview of the RBM algorithm principles.
2. Perform a characteristic analysis of the recommendation data in conjunction with the relevant theories of power-law distribution.

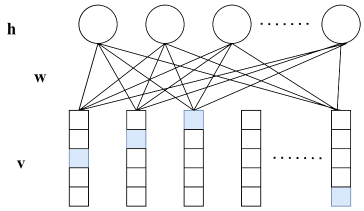


Fig. 3.1: The structure of the RBM Algorithm

3. Modify the Gibbs sampling initial values in the algorithm according to the power-law distribution characteristics of the data. Specifically, in the early stages of iteration, the initial values will be selected randomly, while in the later stages, the initial values will be the results of the previous sampling step.

4. Change the fixed number of Gibbs sampling steps to Periodic Gibbs Sampling (PGS). These improvements aim to offer a reference for enhancing the accuracy of the RBM recommendation algorithm.

3.1. Improvement of Initial Values in Gibbs Sampling for RBM.

3.1.1. Selection of Initial Values in Gibbs Sampling for RBM. Before improving the selection of initial values in the RBM algorithm, it is necessary to briefly introduce the working principles of the RBM algorithm. RBM (Restricted Boltzmann Machine) is a generative stochastic neural network based on an energy function. It is capable of transferring data through visible and hidden layers, deeply learning the latent features of users and items, making it suitable for recommendation problems. The structure of the RBM algorithm is illustrated in Fig. 3.1.

As shown in figure 3.1, v and h represent the visible and hidden units in the visible layer V and the hidden layer H, respectively. a denotes the biases of the visible units, b denotes the biases of the hidden units, and W represents the weights connecting the visible and hidden layers. In the visible layer, a node x_i is multiplied by a weight $W_{i,j}$,then a bias term b is added. The result is then passed through an activation function σ (the sigmoid function) to produce the output of the node x_i.

The energy function for each unit is:

$$E(\nu,h) = -\sum_i a_i \nu_i - \sum_j b_j h_j - \sum_i \sum_j h_j w_{i,j} \nu_i \tag{3.1}$$

Using this energy function, the joint probability distribution between the visible layer and the hidden layer can be obtained:

$$P(\nu,h) = \frac{1}{Z} e^{-E(\nu,h)} \tag{3.2}$$

In equation 3.2, Z is the normalization function that ensures the sum of probabilities over all possible states of the node set $e^{-E(\nu,h)}$ equals 1.

The units within the visible layer and the hidden layer are mutually independent. With the joint probability distribution defined, we can derive the marginal probability distribution, thereby obtaining the activation probabilities of the nodes in the visible layer and the hidden layer.

The units within the visible layer and the hidden layer are mutually independent. With the joint probability distribution defined, we can derive the marginal probability distribution, thereby obtaining the activation probabilities of the nodes in the visible layer and the hidden layer.

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum_{i=1}^{m} w_{i,j} v_i) \tag{3.3}$$

$$p(v_i = 1 \mid h) = \sigma(a_i + \sum_{j=1}^{n} w_{i,j} h_j) \tag{3.4}$$

In equation 3.3 and 3.4, σ represents the sigmoid function. The phase where the hidden layer is computed based on visible layer data during training is referred to as the Positive phase, while the reverse is termed the Negative phase.

Using the visible layer input data again as the starting point, with K ($K \geq 1$) iterations of Gibbs sampling from the visible layer data known, randomly initialize $w_{i,j}$, randomly select the sampling initial value, iterate between the visible and hidden layers using equation 3.3 and 3.4, loop iterate K times, and stop the iteration. The parameter iteration formula is:

$$\begin{cases} \nabla W_{ij} = P(h_j = 1 \mid v^{(0)}) v_i^{(0)} - P(h_j = 1 \mid v^{(k)}) v_i^{(k)} \\ \nabla a_i = v_i^{(0)} - v_i^{(k)} \\ \nabla b_j = P(h_j = 1 \mid v^{(0)}) - P(h_j = 1 \mid v^{(k)}) \end{cases} \tag{3.5}$$

In equation 3.5, $\nu_i^{(0)}$ denotes the sample value, $\nu_i^{(k)}$ represents the sample value obtained after K sampling steps.

Based on the operational process of the RBM algorithm described above, it is evident that Gibbs sampling is the primary iterative algorithm used for recommendation computation in the RBM algorithm. Specifically, the initial values for Gibbs sampling are randomly selected variables from the sample data, and subsequent Gibbs sampling iterations are also based on these initial values, without including updates from the previous iteration steps.

However, the Yelp dataset, after preprocessing using the GRU model to enhance temporal characteristics, indicates that the improved dataset includes contextual feature information. On the other hand, Gibbs sampling parameter updates only consider the parameter update values from the previous step and do not incorporate earlier parameter updates. Consequently, within a limited number of iterations, it is unable to consider the effective information contained in previous parameters.

Therefore, there is a need to enhance the iterative updating method and utilization of information contained in Gibbs sampling parameters.

3.1.2. Improvement of Gibbs Sampling Initialization in RBM. Before improving the Gibbs sampling initialization in RBM, it is necessary to analyze the recommendation data and then refine the initialization based on its characteristics to enhance the accuracy of the CD algorithm.

Analysis of Recommendation Data Characteristics. When making recommendations, users recommend products to other users directly or indirectly based on social relationships. It is necessary to study the characteristics of recommendation data. Using the Yelp dataset, we examine whether there are inherent patterns such as power-law distribution, whose probability distribution is shown in equation 3.6.

$$p(x) = Cx^{-a} \tag{3.6}$$

Power-law distribution refers to a phenomenon where a small number of key items in any given entity contribute to the majority of outcomes or benefits, while the vast majority of items contribute minimally. If recommendation data exhibits power-law distribution characteristics, it indicates that only a small portion of the data contains substantial information, whereas the majority contains minimal information. Therefore, following the approach proposed by Víctor Navas-Portella et al [22], using maximum likelihood estimation to assess the cumulative degree of networks under power-law is recommended. Specifically, for practical datasets, formula 3.7 is employed to estimate the power-law distribution.

$$\alpha \simeq 1 + n[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min} - 0.5}]^{-1} \tag{3.7}$$

In equation 3.7, $\boldsymbol{\alpha}$ represents the power law exponent, and x_i represents the sample data.

Using the user social information from the Yelp dataset, user a following user b is defined as out-degree, and user a being followed by user b is defined as in-degree. Then, the power-law distribution of the frequency of user following and being followed is judged by maximum likelihood estimation. After obtaining and filtering the Yelp dataset, the power-law distribution results are shown in Fig. 3.2.

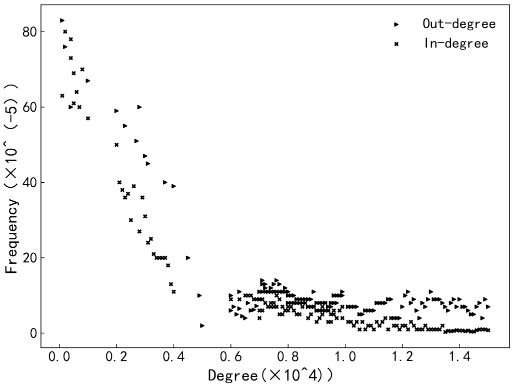


Fig. 3.2: The degree distribution of user attention

From Fig. 3.2, it can be observed that when the number of users is small, there is a higher probability of users following and being followed. As the number of users increases, the frequencies of out-degree and in-degree decrease, and a 'long tail phenomenon' appears at the distribution's tail. This is because when users make hotel choices, the majority of users only follow unfamiliar users who provide more valuable information or reciprocate with friends. Specifically, users with many followers do not necessarily follow all those who follow them. Hence, a small number of users have a high degree of followers, while the probabilities of following and being followed for the majority of users are low. Therefore, the user social information in the Yelp dataset exhibits a power-law distribution.

In summary, the user engagement data in the Yelp recommendation dataset exhibits a long-tail distribution, indicating that a small number of top users have limited social connections, while the majority of users in the tail contribute significantly to the dataset. However, in Gibbs sampling, the random selection of initial values means that if the initial value at time t is sampled from the head of the distribution, the sample at time t+1 could be from either the head or the tail. This randomness across iterations prevents Gibbs sampling from concentrating on gathering data from the tail, thereby limiting the thorough extraction of information from tail-end users.

Similarly, fixed Gibbs sampling steps treat the head and tail segments of the dataset equally, preventing deeper learning from tail-end data. Therefore, when applying the RBM model to predict recommendations from this dataset, it is essential to enhance the parameter iteration and sampling methods of the RBM algorithm to account for the dataset's long-tail characteristics effectively.

Improved Strategy for Gibbs Sampling Initialization in RBM. Due to the long-tail nature of recommendation data, it is evident that the majority of recommendations are concentrated towards the tail end. However, the current method of initializing Gibbs sampling involves randomly selecting training data from the recommendation network. This random selection could pick either head or tail data as initial values, failing to concentrate on tail data and thereby lacking in-depth analysis of this segment, which contradicts the long-tail characteristic of recommendation networks.

Therefore, it is necessary to enhance the strategy for randomly selecting initial values in Gibbs sampling as follows: during the initial training phase, use the original training data as initial values, and during subsequent phases, use the previous Gibbs sampling values as initial values. When the initial value is the original training data, the update method for Gibbs sampling initialization is shown in equation 3.8. When the original data is the previous training data, the Gibbs sampling initialization update method is shown in equation 3.9.

$$\nu^{(0)} \rightarrow \nu^{(k)} \tag{3.8}$$

$$\mathrm{v}^{(k-1)} \rightarrow \nu^{(k)} \tag{3.9}$$

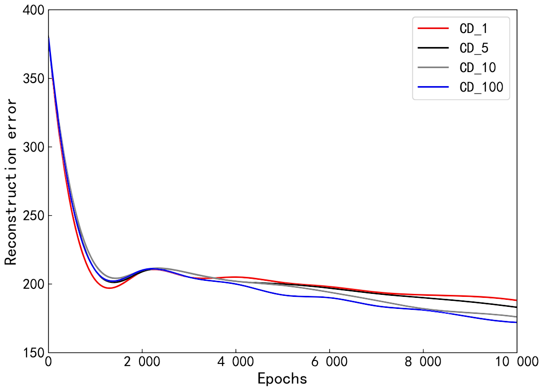


Fig. 3.3: Comparison of CD-K iterative reconstruction error

In equation 3.8 and 3.9, $V^{(0)}$ represents the randomly selected initial training value, and $V^{(k)}$ represents the training value after K steps of Gibbs sampling.

To determine the threshold for changing the Gibbs initial value sampling method, it is necessary to analyze the reconstruction error line chart of CD-K. From Fig.3.3, it can be observed that CD-1, CD-5, CD-10, and CD-100 algorithms show a gradual increase in reconstruction error after approximately 2000 iterations, followed by a slow decrease. This indicates that randomly selecting training data as initial values in the early iterations can lead to rapid convergence of the Gibbs sampling network. However, after about 2000 iterations, the training effectiveness of the RBM model decreases due to slower network convergence. Therefore, the threshold for the number of Gibbs sampling iterations is set at 2000. During iterations 1-2000, random training data is selected as the initial value for sampling, and from 2001 to 10000 iterations, the initial value for sampling is selected as the parameter sampled from the previous Gibbs sampling, ensuring rapid convergence in the early iterations and higher precision convergence in the later iterations.

3.1.3. Analysis of Improvements in Gibbs Sampling Initialization in RBM. Previous sections provided both theoretical and experimental analyses of the characteristics of recommendation data and the improvement strategy for Gibbs sampling initialization in RBM. It was demonstrated that the initial sampling values should be changed from purely random selection to using random values in the early stages of iteration and using the previous step's sampling results in the later stages. This section will validate the effectiveness of the improved Gibbs sampling initialization strategy through parameter gradient verification.

Based on the energy function of the RBM in Equation 3.1 and the marginal probability distributions of the visible and hidden layers in Equations (3.3-3.4), the parameter gradient for iterative parameter updating in the RBM network using Gibbs sampling is given by:

$$\nabla\overset{\wedge}{\theta}_1 = -\sum_h P(h \mid \nu^{(0)} \frac{\partial E(\nu^{(0)}, h)}{\partial \theta}) + E_{P(\nu^{(k)}|\nu^{(0)})}[\sum_h P(h \mid \nu^{(k)} \frac{\partial E(\nu^{(k)}, h)}{\partial \theta})] \tag{3.10}$$

In Equation 3.10,θ represents the general term for the parameters of the RBM.

According to the strategy for improving Gibbs sampling initialization, the updated parameter gradient is given by:

$$\begin{aligned} \nabla\overset{\wedge}{\theta}_2 = &-\sum_h P(h \mid v^{(0)} \frac{\partial E(v^{(0)}, h)}{\partial \theta}) + E_{P(v^{(i)}|v^{(0)})}[\sum_h P(h \mid v^{(i)} \frac{\partial E(v^{(i)}, h)}{\partial \theta})] \\ &-\sum_h P(h \mid \nu^{(i)} \frac{\partial E(\nu^{(i)}, h)}{\partial \theta}) + E_{P(\nu^{(i+1)}|\nu^{(i)})}[\sum_h P(h \mid \nu^{(i+1)} \frac{\partial E(\nu^{(i+1)}, h)}{\partial \theta})] \end{aligned} \tag{3.11}$$

According to Equation 3.11,$-\sum_h P(h \mid \nu^{(0)} \frac{\partial E(\nu^{(0)},h)}{\partial \theta}) + E_{P(\nu^{(i)}|\nu^{(0)})}[\sum_h P(h \mid \nu^{(i)} \frac{\partial E(\nu^{(i)},h)}{\partial \theta})]$represents the Gibbs sampling initialization as the original random sampling data $v^{(0)}$, with i steps of iteration, where i denotes the number of iterations and $0 < i+1 \leq k$.On the other hand, $-\sum_h P(h \mid \nu^{(i)} \frac{\partial E(\nu^{(i)},h)}{\partial \theta}) + E_{P(\nu^{(i+1)}|\nu^{(i)})}[\sum_h P(h \mid \nu^{(i+1)} \frac{\partial E(\nu^{(i+1)},h)}{\partial \theta})]$ represents the Gibbs sampling initialization as the sampling result from the previous step $v^{(i)}$, with $k-1$ steps of iteration. When Gibbs sampling reaches the i-th step, the initialization value is transformed from $v^{(0)}$ to $v^{(i)}$.

The optimization of the RBM model is achieved by finding the optimal parameters through gradient descent.Persistent Contrastive Divergence (PCD) is one of the benchmark algorithms used in RBM training. PCD is proved to be able to approach the network distribution with a small enough learning rate of network parameters. Therefore, using the result of the last parameter iteration as the initial value of the next iteration can make the training parameters change little, and make the parameter gradient decline faster and stabilize in a smaller interval.

Therefore, the parameter gradient of the improved method is smaller than the original randomly selected initial values:

$$\begin{gathered}
-\sum_h P(h \mid \nu^{(i)} \frac{\partial E(\nu^{(i)},h)}{\partial \theta}) + E_{P(\nu^{(i)}|\nu^{(0)})}[\sum_h P(h \mid \nu^{(i)} \frac{\partial E(\nu^{(i)},h)}{\partial \theta})] + E_{P(\nu^{(i+1)}|\nu^{(i)})}[\sum_h P(h \mid \nu^{(i+1)} \frac{\partial E(\nu^{(i+1)},h)}{\partial \theta})] < \\
-\sum_h P(h \mid \nu^{(0)} \frac{\partial E(\nu^{(0)},h)}{\partial \theta}) + E_{P(\nu^{(i)}|\nu^{(0)})}[\sum_h P(h \mid \nu^{(i)} \frac{\partial E(\nu^{(i)},h)}{\partial \theta})] + E_{P(\nu^{(i+1)}|\nu^{(0)})}[\sum_h P(h \mid \nu^{(i+1)} \frac{\partial E(\nu^{(i+1)},h)}{\partial \theta})], i+1 \leq k
\end{gathered} \tag{3.12}$$

In Equation (12),

$$E_{P(\nu^{(i)}|\nu^{(0)})}[\sum_h P(h|\nu^{(i)} \frac{\partial E(\nu^{(i)},h)}{\partial \theta})] + E_{P(\nu^{(i+1)}|\nu^{(0)})}[\sum_h P(h|\nu^{(i+1)} \frac{\partial E(\nu^{(i+1)},h)}{\partial \theta})] = E_{P(\nu^{(k)}|\nu^{(0)})}[\sum_h P(h|\nu^{(k)} \frac{\partial E(\nu^{(k)},h)}{\partial \theta})]$$

That is $\nabla \overset{\wedge}{\theta}_2 < \nabla \overset{\wedge}{\theta}_1$.

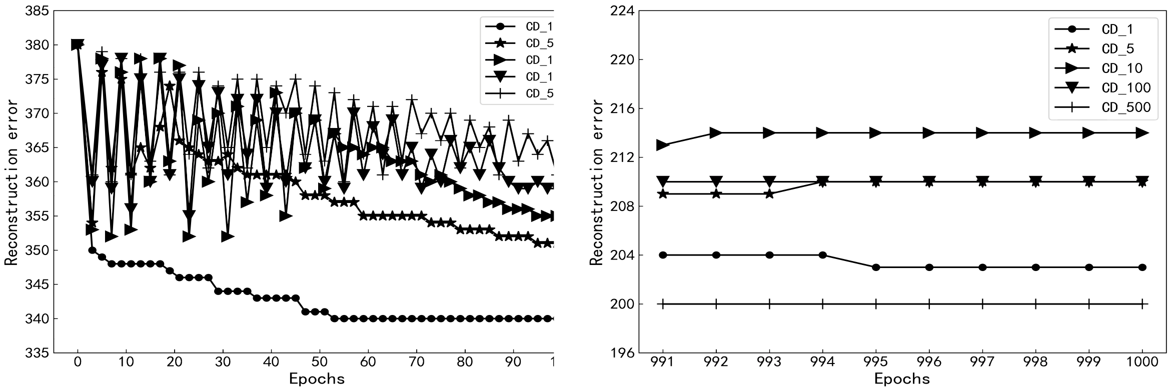
From the perspective of parameter gradients, the improvements to Gibbs sampling initialization are demonstrated to be effective.

3.2. Improving Gibbs Sampling Steps in RBM. Section 3.1.2 analysis highlighted the foundational characteristics of power-law distribution in the Yelp dataset's complex network. However, fixed Gibbs sampling steps collect an equal amount of data from both the head and tail of the dataset, failing to concentrate on learning tail-end information. This results in insufficient characterization of user information in the tail. Therefore, there is a need to improve and adjust the Gibbs sampling steps.

3.2.1. Comparison of Single-step and Multi-step Gibbs Sampling. Section 3.1 has already introduced and analyzed the principles of Gibbs sampling and the long-tail characteristics of recommendation data. Therefore, this section compares single-step Gibbs sampling with multi-step Gibbs sampling to assess their performance advantages and disadvantages. Additionally, leveraging the classical momentum algorithm (CM) to determine decision points for varying Gibbs sampling steps and formulate Gibbs sampling strategies. Finally, from the perspective of Markov chain theory, analyze and justify the rationality of improving Gibbs sampling steps.

Comparison of training errors between single-step Gibbs sampling (CD-1) and various multi-step Gibbs samplings (CD-5, CD-10, CD-100, CD-500) at epochs 1-100 and 991-1000. Utilizing the concept of reconstruction, original data is obtained from trained data, and reconstruction error serves as the evaluation metric to compare CD-K sampling results, thereby assessing the performance of the RBM network at different iteration steps. The comparison results are shown in Fig.3.4.

As shown in Figures 3.4(a) and (b), during the early stages of RBM parameter iteration, the reconstruction error of single-step Gibbs sampling (CD-1) decreases rapidly and vertically, outperforming the training error of multi-step sampling (CD-K, where $(K > 1)$. This indicates that single-step sampling provides better fitting of the training data. In the initial stages of RBM training, with fewer iterations and larger recommendation



(a) Gibbs pre-sampling error

(b) Gibbs post-sampling error

Fig. 3.4: Gibbs sampling error comparison

errors, simple single-step sampling can quickly reduce the recommendation error without the need for time-consuming multi-step sampling. On the other hand, multi-step Gibbs sampling exhibits higher and oscillating reconstruction errors, suggesting larger errors during the early iteration stages. In the later stages of RBM parameter iteration, both single-step and multi-step sampling errors stabilize. However, the training error of single-step sampling is higher compared to multi-step sampling, indicating that random sampling alone can no longer significantly improve recommendation accuracy and emphasizes the need to focus on sampling the tail-end data.

Therefore, Gibbs sampling exhibits strong training capabilities for the RBM model, but its sampling steps significantly impact the algorithm's performance. So the following sections will analyze the influence of Gibbs sampling steps on algorithm performance, combining theoretical analysis with the characteristics of recommendation networks to improve Gibbs sampling.

3.2.2. Improvement Strategies for Gibbs Sampling Steps in RBM. The primary distinction between CD-1 and CD-K lies in their iteration steps, which result in different parameter iteration gradients. Specifically, CD-1 sampling concludes after Gibbs sampling step 1, while CD-K sampling involves K Gibbs sampling steps before terminating the CD algorithm. This leads to the following analysis: CD-1 exhibits good early-stage effectiveness but lacks high precision in later stages, whereas CD-K shows initial error oscillation but achieves higher accuracy in the later stages.

The magnitude of parameter gradients can measure the effectiveness of training methods for parameters. How to divide the sampling steps within the iteration interval can be judged based on the magnitude of gradient ascent to determine the Gibbs sampling steps, thus the classic momentum algorithm (CM) can be used to determine the decision points for Gibbs sampling step changes. CM adjusts the difference between accumulated velocity and current gradient $\nabla g(\theta_t)$ to decrease the target gradient, thereby accelerating the convergence speed of parameter learning. The RBM model is trained based on gradient ascent, hence CM's gradient update formula under the RBM model is shown in equation 3.13 and equation3.14.

$$\nu_{t+1} = \mu\nu_t + \varepsilon\nabla g(\theta_t) \tag{3.13}$$

$$\theta_{t+1} = \theta_t + \nu_{t+1} \tag{3.14}$$

In equation 3.13 and equation 3.14, v_t represents the accumulated velocity, $\nabla g(\theta_t)$ denotes the gradient of the objective function at the current point, θ denotes the parameters of the model, μ represents the accumulated velocity parameter, and ε represents the learning rate.

When performing single-step Gibbs sampling in the initial training stages of the RBM model, as the number of iterations increases, the numerical values of the network weights also increase. As the network weights expand, denoted by $|w| \rightarrow +\infty$, according to equation (3.3-3.4).

$$b_j + \sum_{i=1}^{m} w_{i,j} v_i \rightarrow \infty \tag{3.15}$$

$$a_i + \sum_{j=1}^{n} w_{i,j} h_j \rightarrow \infty \tag{3.16}$$

Therefore, the corresponding probabilities for the hidden layer nodes and visible layer nodes change to:

$$P(h_k = 1 \mid \nu) \rightarrow 0 or 1 \tag{3.17}$$

$$P(\mathrm{v}_k = 1 \mid h) \rightarrow 0 or 1 \tag{3.18}$$

As the number of iterations increases, the sampling probabilities of the Gibbs sampling chain gradually approach 0 or 1. That is, during each sampling, the values at each point are either 0 or 1. At this point, the transition operator of the sampling chain loses its randomness, indicating that the parameter gradient optimization direction is not the fastest. Moreover, the mixing rate of the Gibbs sampling chain decreases as the randomness of its transition operator decreases[23]. This means that as the number of iterations increases and the network weights grow, the mixing rate of the single-step Gibbs sampling chain gradually decreases, leading to reduced accuracy in later stages. Similarly, multi-step Gibbs sampling involves K repetitions of single-step Gibbs sampling, which confirms that multi-step Gibbs sampling may experience slower convergence in the early iterations.

Based on the analysis above and the results in Fig. 3.3, it is evident that improving the parameter iteration method of the RBM model in conjunction with the characteristics of the recommendation data can enhance the efficiency of the recommendation algorithm[24, 25]. Specifically, reducing the number of sampling steps in the sparse head of the data and increasing the sampling steps in the tail can yield more accurate data information.

Comparing the Gibbs sampling of CD-1, CD-5, CD-10, and CD-100 as shown in Fig.3.3, CD-1 exhibits better reconstruction error during the early iterations (1-2000 iterations) compared to multi-step Gibbs sampling. However, beyond this range, CD-10 consistently outperforms other step sizes in sampling.Therefore, the strategy for improving Gibbs sampling steps is as follows.

1. For iterations 1 to 2000, set Gibbs sampling steps $\mathrm{K}_1 = 1$. Execute single-step Gibbs sampling, using equation3.3 and equation 3.4, to compute the probability distributions of visible and hidden layers.
2. For iterations 2001 to 10000, set Gibbs sampling steps $\mathrm{K}_2 = 5$. Execute 5-step Gibbs sampling, using equation3.3 and equation 3.4, to compute the probability distributions of visible and hidden layers.

3.2.3. Analysis of the Improvement Characteristics of Gibbs Sampling Steps in RBM. Gibbs sampling is a type of Markov Chain Monte Carlo (MCMC) sampling algorithm. This section analyzes the number of Gibbs sampling steps using relevant theories from Markov chains, providing a theoretical justification for the improvement in the number of sampling steps.

In RBM model training, hidden layer nodes and input layer nodes are sampled alternately, as described in Equations 3.3 and 3.4. According to the Markov chain convergence theorem, if the number of possible states for the parameters is finite, the transition probabilities of the chain are fixed, and the parameter states can transition from any state to any other state. Therefore, when the number of steps $n \rightarrow +\infty$, the Gibbs sampling chain will converge to a stationary distribution:

$$\pi_i(x) = \pi_{i-1}(x)P = \pi_0 P^n, i \in S \tag{3.19}$$

In Equation 3.19, $\pi_i(x)$ represents the stationary distribution of the sample x. i denotes an arbitrary state, and S is the state space.

Furthermore, according to the detailed balance criterion of Markov chains, it can obtain:

$$\pi(x_i)P_{ij} = \pi(x_j)P_{ji}, \forall i, j \in S \tag{3.20}$$

In Equation 3.20,x_i and x_j represent the training data. P_{ij} and P_{ji} denote the Markov transition probabilities.

According to Equation 3.20, the stationary distribution achieved by Gibbs sampling is independent of the initial sampling values and depends only on the Markov transition probabilities. Combining this with the alternating sampling probability formulas for the visible and hidden layers in the RBM algorithm (Equations 3.3 and 3.4), it can be seen that when using Gibbs sampling for iterative training of RBM model parameters, the stationary distribution is a function of the network parameters:

$$\pi(x) = f(a, b, w) \tag{3.21}$$

In Equation 3.21, $\pi(x)$represents the stationary distribution of the sample x. $f(\theta)$ denotes the distribution of the parameters as a function.

The trained parameter values are denoted as $\hat{\theta} = (\hat{a}, \hat{b}, \hat{w})$, while the true parameter values are $\theta = (a, b, w)$. The goal of training the RBM is to adjust the network parameters such that the trained parameter values are as close as possible to the true parameter values.

$$\begin{cases} \Delta a = \hat{a} - a \\ \\ \Delta b = \hat{b} - b \\ \\ \Delta w = \hat{w} - w \end{cases} \tag{3.22}$$

Therefore, in the early stages of RBM iterative learning, when the trained parameter values are significantly different from the true values, multiple steps of Gibbs sampling can cause the sampled values to deviate further from the true values, resulting in multiple oscillations in the parameter values during early iterations. Single-step sampling, however, allows for faster convergence of the parameters to the true values. In the later stages of sampling, as the trained parameter values approach the true values, multiple-step sampling can more deeply capture the latent features of the recommendation data, thereby improving the accuracy of parameter training. On the other hand, single-step sampling may lead to a path with significant deviation from the true values, making CD-1 susceptible to local minima and resulting in lower parameter accuracy in the later stages of single-step sampling.

From the above analysis, it is evident that both single-step Gibbs sampling and multi-step Gibbs sampling have limitations in the RBM network training process, as demonstrated by MCMC algorithms. The experimental results, as shown in Figure 3.4, further validate this from the perspective of Markov chains. This highlights the necessity of changing the fixed sampling step size in Gibbs sampling to a phase-based variable step size.

4. The whole process of improving RBM algorithm is introduced. The improvements made to the RBM model itself primarily focus on refining the initial values and sampling steps of Gibbs sampling. Specifically, the random selection of Gibbs sampling initial values has been adjusted to a combination of random selection and the previous sampling value, and the fixed Gibbs sampling steps have been changed to staged Gibbs sampling steps. The specific improvement process is as follows: First, determine a model iteration of 10,000 steps. Then, during iterations 1-2,000, initialize sampling with randomly selected training data from the recommendation set and set the sampling step to 1. For iterations 2,001-10,000, use the previous Gibbs sampling result as the sampling value and set the sampling step to 5. The algorithm process is illustrated in Algorithm 1.

Combined with the Data pre-processing analysis, the overall process of the recommender system improvement is shown in Fig.4.1.

Algorithm 1

```
Step 1: Set initial values for RBM model parameters.
Step 2: Input the Yelp dataset into the visible layer of the RBM model, randomly select initial sampling values,
        and adjust the fixed sampling steps to staged sampling steps, setting the staged sampling step to 1.
Step 3: Within iterations 1-2000, perform interactive sampling between the RBM visible and hidden layers
        using equations 3.3 and 3.4, and update parameters using equation 3.5 with the sampled values.
Step 4: Repeat steps 2-3 for 2000 times to complete single-step Gibbs sampling with a sampling step of 1.
Step 5: Within iterations 2001-10000, change the randomly selected initial sampling value to the previous
        sampling value, and adjust the 1-step Gibbs sampling step from step 2 to 5.
Step 6: Repeat steps 3 and 5 for 8000 times. Stop parameter training at iteration 10000, completing Gibbs
        sampling with a step of 5.
```

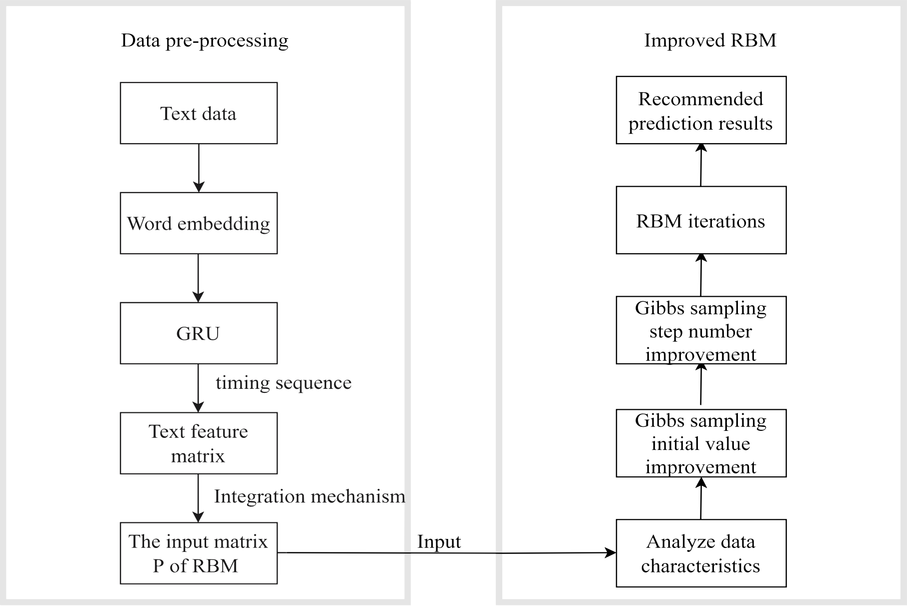

Data pre-processing
Text data
Word embedding
GRU
timing sequence
Text feature matrix
Integration mechanism
The input matrix P of RBM
Input
Improved RBM
Recommended prediction results
RBM iterations
Gibbs sampling step number improvement
Gibbs sampling initial value improvement
Analyze data characteristics

Fig. 4.1: Personalize the recommendation process

5. Experimental analysis. The previous section analyzed the power-law characteristics of the Yelp dataset and improved the Gibbs sampling's random initial values and the number of sampling steps based on the characteristics of the recommendation data. The characteristics of the improved initial values and sampling steps were analyzed, and the effectiveness of the improved RBM algorithm was discussed from a theoretical perspective. Therefore, this section aims to conduct an empirical analysis of the Gibbs sampling and RBM recommendation algorithms before and after the improvements, to further explore the effectiveness of the improved RBM recommendation algorithm.

The dataset selected for the empirical analysis is the improved Yelp dataset, which has been refined according to the data preprocessing steps shown on the left side of Figure 4.1. This includes GRU sequential processing of the textual information in the dataset, quantifying the different contributions of the text data using the attention mechanism, and integrating the text data with the rating data based on user preferences. The details of the improved Yelp dataset are shown in Table5.1.

Compare the reconstruction error metrics of the initial values and sampling steps of Gibbs sampling before and after the improvements. Additionally, analyze the recall@K, MAE, and RMSE metrics using classical recommendation algorithms such as RBM, RTRBM, PMF, and the improved RBM algorithm. Finally, analyze the experimental results to determine the effectiveness of the improved RBM recommendation algorithm strategy.

Table 5.1: Features of the improved Yelp data set

Features	Improved Yelp dataset
number of users	15328
number of items	37335
number of ratings	106821
number of comments	13796
number of mutual attention	4626

5.1. Introduction to the Dataset, Metrics, and Comparison Algorithms.

5.1.1. Introduction to the Dataset. The Yelp dataset is an online service and business review platform where users can post reviews and ratings for businesses. It includes a wealth of data such as user reviews, ratings, user information, and business information. The details of the improved Yelp dataset are shown in Table 5.1.

5.1.2. Introduction to Metrics. The effectiveness of the improvements to the RBM algorithm is measured using the following three metrics. recall@K indicates the proportion of correctly predicted positive samples out of all positive samples.

$$recall@k = \frac{TP@k}{TP@k + FN@k} \tag{5.1}$$

In Equation 5.1,$TP@k$ represents the number of correctly recommended items in the top-K recommendation list, and $FN@k$ represents the number of incorrectly recommended items in the top-K recommendation list. A higher recall@K value indicates better performance of the model's recommendations.

MAE stands for Mean Absolute Error, and RMSE stands for Root Mean Squared Error. Both metrics are used to measure the difference between predicted values and observed values in recommendation algorithms.

$$MAE = \frac{1}{T} \sum_{(u,i) \in T} \mid r_{ui} - r_{ui}^{\wedge} \mid \tag{5.2}$$

$$RMSE = \sqrt{\frac{1}{T} \sum_{(u,i) \in T} (r_{ui} - r_{ui}^{\wedge})^2} \tag{5.3}$$

In Equations 5.2 and 5.3, T represents the test set, r_{ui} and $r_{ui}^{\wedge}$ denote the true rating and the predicted rating for item by user, respectively. Generally, smaller MAE and RMSE values on the test set indicate better accuracy in the rating predictions of the recommendation algorithm.

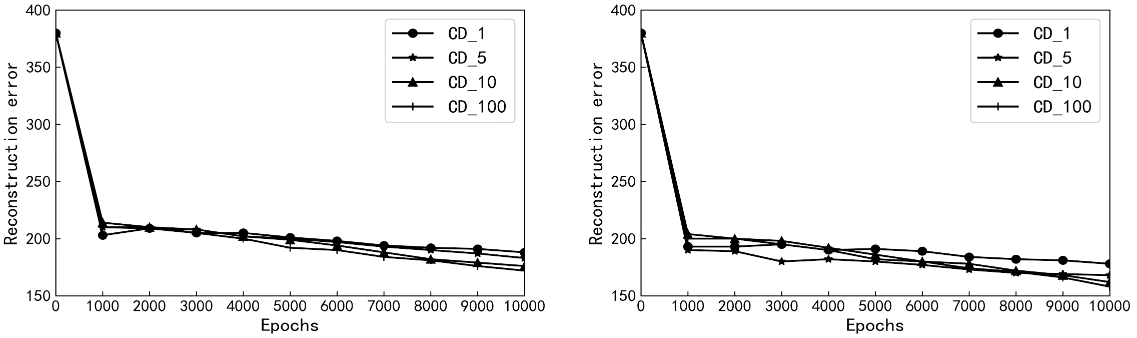
5.1.3. Introduction to Comparison Algorithms.

(1) Restricted Boltzmann Machine (RBM) Algorithm. Using the Yelp recommendation dataset, RBM processes data through visible and hidden layers, employing Gibbs sampling to deeply learn the latent features of users' hotel choices, thereby generating hotel recommendation predictions.

(2) Improved Restricted Boltzmann Machine (RBM) Algorithm. Building upon RBM, this improved version adjusts the sampling initial values and sampling steps of Gibbs sampling to phased sampling initial values and phased sampling steps.

(3) Recurrent Temporal Restricted Boltzmann Machine (RTRBM) Algorithm. This algorithm can be seen as an extension of RBM, where several RBMs are horizontally concatenated. It utilizes continuously sampled differences from previous sampling results to effectively handle temporal information about changes in user preferences for hotels.

(4) Probabilistic Matrix Factorization (PMF) Algorithm. This algorithm leverages both user social information and user rating and review information. It decomposes the "user-social attention" matrix into user implicit factor matrices and social attention implicit factor matrices. Using these implicit factor matrices, it predicts user ratings for hotels and generates hotel recommendation lists.



(a) Reconstruction error of Gibbs sampling initial value before improvement

(b) Reconstruction error of improved Gibbs sampling initial value

Fig. 5.1: Comparison of reconstruction error of Gibbs sampling random initial value before and after improvement

5.2. Comparative analysis of Gibbs sampling initial values. By comparing the random initial values of Gibbs sampling with the initial values set as random for the early iterations (1-2000 iterations) and as the values from the previous Gibbs sampling for the later iterations (2001-10000 iterations), the effectiveness of the improved Gibbs sampling initial values can be demonstrated. Therefore, under the premise of 10,000 iterations, the reconstruction error metrics of the initial values of Gibbs sampling before and after the improvements for CD-1, CD-5, CD-10, and CD-100 are calculated to determine the effectiveness of the improved Gibbs sampling initial values strategy. The experimental results are shown in Fig.5.1.

1. As shown in Fig.5.1(a), the selection of random initial values for Gibbs sampling does not satisfy the power-law distribution characteristics of the recommendation data. This results in higher reconstruction errors for both single-step and multi-step sampling during the 10,000 iterations compared to the improved initial values of Gibbs sampling shown in Fig.5.1(b). This demonstrates that the improvement to the Gibbs sampling initial values—using random initial values in the early iterations (1-2000 iterations) and the previous sampling results in the later iterations (2001-10000 iterations)—enables the early iterations to converge quickly and the later iterations to converge to higher precision, effectively reducing the reconstruction error throughout the entire training cycle.
2. Meanwhile, the change in CD-1 reconstruction error before and after the improvement is minimal, decreasing from 188 to 178 in the later stages. This is because single-step sampling converges extremely quickly within just one iteration, resulting in a slower decrease in reconstruction error. However, the greater the number of multi-step samplings, the larger the reduction in reconstruction error. This indicates that multi-step sampling focuses on collecting tail data in the later stages, allowing the tail data of the recommendation data to be intensively learned through iterations, thereby improving the accuracy of Gibbs sampling.
3. Considering the characteristics of the Yelp dataset, users' ratings and reviews of hotels are quite sparse, with a sparsity level exceeding 90%. Only a few users who frequently patronize hotels and actively review them receive a lot of attention. When the RBM recommendation algorithm learns the users' rating and review information for hotels, it processes all user information one by one, leading to the loss of hotel review data for the majority of users. This results in a decrease in the recommendation accuracy of RBM. Therefore, by changing the strategy of random initial values in Gibbs sampling and shifting the algorithm's attention to users with more social information and hotel ratings and reviews, the reconstruction error of Gibbs sampling can be reduced, thereby enhancing the predictive performance of the RBM recommendation algorithm.

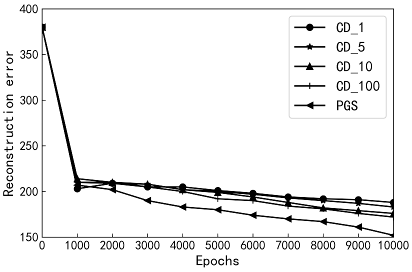


Fig. 5.2: Comparison of reconstruction error between fixed Gibbs sampling steps and staged Gibbs sampling steps

In summary, after improving the initial values of Gibbs sampling, the RBM algorithm can adjust sampling based on the characteristics of the concentrated tail data in the recommendation dataset, making the advantages of the improved RBM algorithm more apparent. The improved RBM algorithm is suitable for large-scale and sparse recommendation data, enhancing the learning speed of the algorithm while ensuring the accuracy of the recommendations.

5.3. Comparative analysis of Gibbs sampling steps. Based on the empirical analysis of the Gibbs sampling initial values mentioned above, a similar empirical comparative analysis is conducted for the Gibbs sampling steps before and after the improvements. By comparing the performance of fixed Gibbs sampling steps with the strategy of single-step sampling in the early iterations (1-2000 iterations) and multi-step sampling (CD-5) in the later iterations (2001-10000 iterations), with the initial values being random in both cases, we aim to validate the effectiveness of phased sampling steps. Therefore, under the premise of 10,000 iterations, the reconstruction error of fixed Gibbs sampling steps CD-1, CD-5, CD-10, CD-100, and phased sampling steps is calculated to determine the effectiveness of the phased Gibbs sampling strategy. The experimental results are shown in Fig.5.2.

1. As shown in Fig.5.2, during the early iterations, the phased Gibbs sampling (PGS) employs single-step sampling, resulting in significant decreases in reconstruction errors for various fixed Gibbs sampling steps (CD-K) and PGS. This rapid convergence indicates comparable performance across these methods. In the later iterations, the reconstruction error of PGS is lower than that of fixed Gibbs sampling steps. This suggests that PGS facilitates rapid convergence in the early iterations while focusing on improving sampling precision in the later iterations, effectively enhancing the operational efficiency and accuracy of the RBM algorithm. Furthermore, it demonstrates PGS's ability to adapt effectively to the power-law distribution characteristics of recommendation data by conducting multi-step sampling in the tail data, thereby improving the accuracy of Gibbs sampling.
2. According to the analysis of the reconstruction error characteristics of Gibbs sampling before improvement, fixed Gibbs sampling steps do not conform to the power-law characteristics of the Yelp dataset. From the reconstruction error characteristics of PGS shown in Fig.5.2, it is evident that throughout the 10,000 iterations, PGS consistently maintains lower reconstruction errors compared to fixed sampling steps. This indicates that PGS can effectively leverage the rich information of users in the Yelp dataset, enabling better selection and prediction of hotels.
3. During the initial 2000 iterations, whether using single-step or multi-step Gibbs sampling, the reconstruction error shows a significant decrease followed by a slight increase. This pattern is determined by the characteristics of the Yelp dataset. In the early stages of recommendation prediction, the Yelp dataset allows RBM to learn and predict recommendations based on limited information. This initial learning involves deep learning of relevant information from scratch, acquiring data features, and thus achieving a significant decrease in reconstruction error after a few Gibbs iterations. However, the majority of users in the dataset have sparse information, providing minimal evaluations on aspects

Table 5.2: Comparison of performance indicators of different recommendation algorithms

Recommendation algorithm Indicators	RBM		**Improved RBM**		RTRBM		PMF	
recall@K	0.728		**0.763**		0.731		0.712	
MAE	0.801	0.833	0.741	0.702	0.778	0.764	0.862	0.848
RMSE	1.291	1.227	1.114	1.031	1.201	1.126	1.429	1.354

such as hotel location and cuisine. Due to the drawbacks of fixed single-step and multi-step Gibbs sampling namely, insufficient precision in the later stages and high time costs respectively the original RBM algorithm struggles to efficiently predict recommendations based on the characteristics of the Yelp dataset.

In conclusion, throughout the entire iteration cycle, PGS demonstrates its ability to balance the parameter iteration of the RBM algorithm, allowing the network to converge to a higher precision. Specifically, in the early stages of the algorithm's recommendation learning, single-step Gibbs sampling facilitates faster parameter convergence, enabling the parameters to approach the true values. In the later stages, rapid parameter convergence through multi-step sampling also helps the parameters to converge to the true values. This indicates that the improvement in Gibbs sampling steps is meaningful and can effectively enhance the recommendation efficiency of the RBM.

5.4. Performance Comparison Analysis of Different Recommendation Algorithms. In addition to comparing the initial values and sampling steps of Gibbs sampling before and after improvement, it is also essential to conduct a comparative analysis between the improved RBM algorithm and other classical recommendation algorithms to assess the performance of the improved RBM. This section will compare the original RBM algorithm, Recurrent Temporal Restricted Boltzmann Machine (RTRBM) algorithm, Probabilistic Matrix Factorization (PMF) algorithm, and the improved RBM algorithm based on recall@K, MAE, and RMSE metrics. This comparison aims to demonstrate the effectiveness of the improved RBM algorithm.

5.4.1. The results of the experiment. A comparison was made between the classic recommendation algorithms and the improved RBM. The performance of the algorithms was evaluated using metrics such as recall@K, MAE, and RMSE. The performance comparison results for different recommendation algorithms are shown in Table 5.2. The performance of the algorithms was tested using different test set sizes, with the test sets being 70% and 80%, and the prediction sets being 30% and 20%. In Table 5.2, the MAE and RMSE values on the left side correspond to the predictions with the 70% test set, while those on the right side correspond to the predictions with the 80% test set.

5.4.2. Analysis of experimental results.

1. As shown in Table 5.2, the recall@K value of the improved RBM algorithm is 0.763, which is higher than that of the other algorithms. In terms of MAE and RMSE, when the test set is 70%, the improved RBM algorithm yields values of 0.741 and 1.114, respectively, which are lower than those of the other algorithms. When the data sparsity is further reduced, the improved RBM algorithm values decrease to 0.702 and 1.031. This clearly demonstrates that the improvements made to the sampling initial values and sampling steps of the RBM algorithm, considering the power-law characteristics of the dataset, effectively enhance both the efficiency and accuracy of recommendation predictions.
2. All metrics in Table 5.2 demonstrate that the performance of the improved RBM algorithm surpasses that of the original RBM algorithm. This indicates that compared to the original RBM algorithm, the improved RBM algorithm can deeply learn various aspects of information from the Yelp dataset, including user preferences for hotel location, cuisine, service, and social interactions, thereby achieving superior recommendation performance. During the early stages of algorithm iteration, RBM recommendation involves the transmission and updating of data between visible and hidden layers to learn

relevant features based on user preferences for hotel selection. Therefore, when randomly selecting recommendation data during Gibbs sampling to gather diverse user hotel preference information and mitigate the limitations of single sampling, it effectively reduces algorithm runtime and enhances recommendation efficiency.Similarly, in the later stages of iteration, due to the continuity of user information, the RBM algorithm needs to learn temporal user characteristic information, acquiring hotel consumption information and characteristics over time to better understand changes in user hotel preferences. Hence, considering the temporal aspects and refining the interpretation of user preference features during Gibbs sampling can effectively improve algorithm performance.

3. The recommendation prediction performance of the RTRBM algorithm lies between that of the improved RBM algorithm and the RBM algorithm. The recall@K value of RTRBM is 0.731, which is greater than the values of 0.728 for RBM and 0.712 for PMF. For both the 70% and 80% test sets, the MAE and RMSE values are 0.764 and 1.126, respectively, both of which are higher than the values of 0.833 and 1.227 for RBM, and 0.848 and 1.345 for PMF. This suggests that user preferences for hotels in the Yelp dataset change over time, showing temporal dynamics where preferences in one period influence those in subsequent periods. RTRBM demonstrates efficient capabilities in collecting and organizing user feature information and capturing temporal changes in hotel preferences. This enables it to make highly accurate recommendation predictions within a relatively short timeframe.
4. The recommendation prediction performance of the PMF algorithm is relatively poor. the recall@K, MAE, and RMSE values are 0.712, 0.848 and 1.354, respectively when the test set is 80%.This could be attributed to the PMF algorithm's reliance solely on implicit factor matrices to predict user hotel preferences, without adequately addressing the temporal dynamics of user data. Furthermore, influenced by the sparsity of the dataset, the matrix sparsity in PMF leads to decreased accuracy in prediction results.

In summary, regardless of which metric is used to assess recommendation accuracy, the improved RBM algorithm consistently outperforms others. This demonstrates that the improved RBM algorithm, which considers the characteristics of the input dataset, achieves the best recommendation performance.

6. Conclusion. Most of the studies do not consider the characteristics of real data sets when making recommendation prediction, so this paper studies the power-law distribution characteristics of recommendation data, according to this characteristic, a novel recommendation and prediction algorithm based on improved RBM model is proposed.

According to the long tail characteristic of the recommendation data, the recommendation algorithm is required to collect the recommendation tail data and to study and analyze the tail data deeply. Therefore, the main algorithm in RBM, Gibbs sampling, has been modified: random sampling for initial stages and using the previous sampling results as initial values in later stages, alongside phased sampling steps. This approach aims to concentrate on collecting data from the tail end of recommendations, iteratively analyzing this data to enhance algorithm performance.

Subsequently, the improved Yelp dataset is selected as the training data for the RBM algorithm, and ablation experiments are conducted on Gibbs sampling. The improved RBM is then compared and analyzed against the original RBM, RTRBM, and PMF algorithms. Experimental results demonstrate that the improved RBM algorithm outperforms the other three algorithms in prediction accuracy. It accurately predicts user hotel preferences, effectively enhancing the recommendation prediction capability of the RBM algorithm.

7. Acknowledgment. Natural Science Foundation of Gansu Province: Research on the Evolution of Power Law Distribution of Decentralized Social Networks (Project No. 23JRRA898)

REFERENCES

[1] Garima Gupta, Rahul Katarya., *A computational approach towards food-wine recommendations[J].* , Expert Systems with Applications. 2024, 238(15):121766.

[2] Ralph Bulanadi and Patrycja Paruch., *Outperforming RBM Feature-Extraction Capabilities by "Dreaming" Mechanism[J]. IEEE Transactions on Neural Networks and Learning Systems*, Expert Systems with Applications., 2024, 238(15):121766.

[3] Fang Yang, Tan Zhen, Chen Ziyang, et al., *Meta-learning of heterogeneous information networks for cold start recommendations[J]*, Journal of Software, 2023, 34(10): 4548-4564.

[4] Wang Honglin, Yang Dan, Nie Tiezheng, et al., *A computational approach towards food-wine recommendations[J]*, Computer Research and development,2022, 59(07): 1509-1521.

[5] Ambikesh, G., Rao, S.S. & Chandrasekaran, K., *A grasshopper optimization algorithm-based movie recommender system[J].* , Multimed Tools And Applications., 2024, 83(7), 54189-54210 .

[6] Hafez MM, Redondo RPD, Vilas AF, Pazó HO., *Multi-Criteria Recommendation Systems to Foster Online Grocery[J].* , Expert Systems with Applications., Sensors. 2021, 21(11):3747.

[7] Jha, G.K, Gaur, M, Ranjan, P, et al. , *A trustworthy model of recommender system using hyper-tuned restricted boltzmann machine[J]*, Multimed Tools and Applications., 2022, 82(2): 8261–8285.

[8] GM Harshvardhan, Mahendra Kumar Gourisaria, Siddharth Swarup Rautaray, et al., *UBMTR: Unsupervised Boltzmann machine-based time-aware recommendation system[J]*, Journal of King Saud University - Computer and Information Sciences., 2022, 34(8): 6400-6413.

[9] A. Fachechi, A. Barra, E. Agliari and F. Alemanno(delete) ,et al., *Outperforming RBM Feature-Extraction Capabilities by "Dreaming" Mechanism[J]*, IEEE Transactions on Neural Networks and Learning Systems,2022, 4(34): 1-10.

[10] Xie Miao, Deng Yulin, Lv Jie., *Personalized recommendation algorithm based on depth Restricted Boltzmann machine[J]*, Data acquisition and processing, 2022, 37(02): 456-462.

[11] Jue Wu, Lei Yang, Fujun Yang, et al., *Hybrid recommendation algorithm based on real-valued RBM and CNN[J]*,Mathematical Biosciences and Engineering, 2022, 19(10): 10673-10686.

[12] Tieleman T., *Training restricted Bo;tzmann machines using approximations to the likelihood gradient[C]// Proceedings of the 25th International Conference on Machine Learning*, New York: ACM,2008: 1064-1071.

[13] Tieleman T, Hinton G. Montreal, , *Using fast weights to improve persistent contrastive divergence[C]// Proceedings of the 26th International Conference on Machine Learning (ICML)* , Quebec, Canada: ACM, 2009: 1033-1040.

[14] Li Fei, Gao Xiaoguang, Wan Kai Fang. , *Research on RBM training algorithm based on dynamic Gibbs Sampling[J]*, ACTA automatica, 2016, 42(06): 931-942.

[15] Ramasamy Savitha, ArulMurugan Ambikapathi, Kanagasabai Rajaraman. , *Online RBM: Growing Restricted Boltzmann Machine on the fly for unsupervised representation[J]*, Applied Soft Computing, 2020, 92(12): 422-450.

[16] Li Fei, GAO Xiaoguang and WAN Kaifang., *Training Restricted Boltzmann Machine Using Gradient Fixing Based Algorithm[J]*, Chinese Journal of Electronics, 2018, 27(4): 694-703.

[17] Ma X, Wang X. , *Average Contrastive Divergence for Training Restricted Boltzmann Machines[J]*, Entropy, 2016; 18(1):35.

[18] Kirubahari, R., Amali, S.M.J., *An improved restricted Boltzmann Machine using Bayesian Optimization for Recommender Systems[J]*, Evolving Systems, 2023, 21(66): 34-45.

[19] [Wang, Q., Gao, X., Li, X. et al. , *A precise method for RBMs training using phased curricula[J]*, Multimed Tools Appl, 2022,82(3): 8013–8047.

[20] He Ying,Wang Zhuoran, Zhou Xu,et al. , *Point of interest recommendation algorithm based on weighted matrix decomposition of social geographic information[J]*, Journal of Jilin University Science (engineering edition),2023,53(09):2632-2639.

[21] Shao Changcheng, Chen Pinghua., *Points of interest recommendations that incorporate social networking and graphic content[J]*, Computer Applications, 2019, 39(05): 1261-1268.

[22] Víctor Navas-Portella, Álvaro González, Isabel Serra, et al., *Universality of power-law exponents by means of maximum-likelihood estimation[J]*,Physical Review,2019, 100(6): 062106.

[23] Bengio Y, Delalleau O., *Justifying and generalizing contrastive divergence[J]*, Neural Computation, 2009, 21(6): 1601-1621.

[24] Zhang Shusen, Wei Yudang, Liang Xun, et al. , *Power-law distribution of mobile social networks and identification of kinship[J]*, Chinese Journal of information, 2018, 32(06): 114-123.

[25] Yin H, Cui B, Li J, et al., *Challenging the long tail recommendation[J]*, Proceedings of the VLDB Endowment, 2012, 5(9): 896-907.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Jun 24, 2024
Accepted: Oct 15, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1035–1047, DOI 10.12694/scpe.v26i3.4199

TASK OFFLOADING AND COLLABORATIVE BACKHAUL SYSTEM BASED ON MULTI-LEVEL EDGE COMPUTING IN THE INTERNET OF VEHICLES

JIN LIN[*], ZEQIN LI[†], RUOFEI WANG[‡], RUYUE GONG[§] AND HONGJING WU[¶]

Abstract. With the development of 5G and the Internet of Vehicles, diverse in-vehicle services continue to emerge. Computation-intensive and delay-sensitive in-vehicle tasks pose significant challenges to in-vehicle devices and represent one of the bottlenecks limiting the development of Internet of Vehicles technology. This paper proposes a Speed-Sensitive Offloading (SSO) and collaborative backhaul solution to address the problem of task offloading and result backhaul failure caused by vehicle movement, including a multi-level MEC architecture solution, speed sensitive task offloading and an MEC collaboration-based task return scheme (SSCOM). Through preliminary experimental verification, as the number of vehicles increases, the average task offloading time of all schemes shows an upward trend, but the SSCOM scheme has the smallest increase; compared with the schemes of random offloading, speed-prioritized and data-volume-prioritized offloading, the present scheme can signifi- cantly reduce the average task offloading time; collaborative backhaul can also solve the problem of result backhaul failure caused by vehicles driving out of the coverage area, etc., can improve the task backhaul success rate and MEC resource utilization rate by at least 5%.

Key words: Multi-level Edge Computing, Vehicle, Offloading, Backhaul, Velocity Sensing

1. Introduction.

Currently, 5G and vehicular networking technologies are rapidly advancing, enabling mobile vehicles to interconnect with networks, artificial intelligence, and other intelligent terminal devices. At the same time, various in-vehicle services continue to emerge, such as online navigation, road condition recognition, audio-visual entertainment, and autonomous driving [2, 14]. Although 5G communication technology has improved data transmission speeds, the high mobility and dynamism of vehicles pose significant challenges to in-vehicle devices for computationally intensive, latency-sensitive tasks such as task offloading and collaborative backhaul [8]. An effective solution is to offload computational tasks to cloud servers[11]. However, for delay-sensitive tasks, the delay caused by inter-cloud transfers is almost unacceptable. For example, the delay requirement for task response in autonomous driving is only 5-10ms [5]. In addition, a large number of in-vehicle tasks sending data to the cloud at the same time can also put a huge pressure on the bandwidth of the core network [6].

The emergence of Multi-Access Edge Computing (MEC) [16] provides the possibility to solve this challenge. MEC technology sinks the computing power of the cloud to the edge, and takes advantage of the edge's close proximity to the end-users to effectively reduce the transmission delays of tasks and optimize the user experience [4]. In the application scenario of Internet of Vehicles, MEC is deployed at the edge side of the network such as Road Side Unit (RSU). Therefore, users can offload computational tasks to MEC servers through Vehicle to Infrastructure (V2I) communication, thus relieving the computational pressure on in-vehicle devices, reducing the power consumption of in-vehicle devices, and the transmission delay of tasks [18]. On the other hand, RSUs can also obtain the position, speed and other information of vehicles within the coverage area through V2I communication [15], which provides the possibility for optimal scheduling of vehicle resources.

This paper proceeds to delve into the design and evaluation of a dynamically adaptive multi-level MEC framework. It initiates by addressing task offloading optimization under vehicle mobility constraints and introduces a speed-sensitive offloading strategy. Following this, a collaborative backhaul system is proposed to

[*] Neusoft Institute Guangdong, School of Computer Science, Foshan ,Guangdong, China (linjin@nuit.edu.cn)

[†] Neusoft Institute Guangdong, School of Computer Science, Foshan, Guangdong, China (leezeqin@163.com)

[‡] Neusoft Institute Guangdong, School of Computer Science, Foshan, Guangdong, China (Corresponding author, wangruofei@nuit.edu.cn)

[§] Neusoft Institute Guangdong, School of Computer Science, Foshan, Guangdong, China (gongruyue@nuit.edu.cn)

[¶] Neusoft Institute Guangdong, School of Computer Science, Foshan, Guangdong, China (wuhongjing@nuit.edu.cn)

tackle issues related to task result returns when vehicles exit RSU coverage zones. The paper further presents an improved genetic algorithm tailored for task prioritization and scheduling, and evaluates its efficacy through rigorous experimentation. In conclusion, the paper delves into the intricacies of a pivotal real-time position updating mechanism integrated within the MEC servers. This mechanism is designed to significantly bolster the system's responsiveness and adaptability. It operates by dynamically synchronizing the geographic coordinates of vehicles with the multi-level MEC architecture, ensuring that the computational offloading and result backhaul processes are aligned with the vehicles' instantaneous positions. The seamless interaction with the MEC hierarchy is facilitated through a suite of algorithms and protocols that are adept at handling the velocity and mobility patterns of vehicles. These include, but are not limited to, geo-fencing algorithms for defining coverage areas, vehicle mobility prediction models that forecast short-term movement based on historical data, and real-time communication protocols that enable efficient information exchange between the vehicle and MEC nodes.

2. Related Works. The high mobility and dynamics of vehicular networks pose significant operational challenges, especially in the context of 5G and the Internet of Things (IoT). In these environments, the fast movement of vehicles leads to frequent switching and coverage transitions, which affects the stability and efficiency of task offloading and collaborative backhaul. Recent research has highlighted the importance of addressing these challenges. Sheng et al. [12] demonstrated that in edge sensor networks, tasks can be divided into subtasks for distributed processing. However, their work did not adequately consider the impact of vehicle mobility on the effectiveness of task offloading. Similarly. Chen et al. [3],You and Huang [17] focused on the division of computational tasks but did not sufficiently address the challenge posed by the movement of vehicles. Zhou et al. [20] developed a model for task segmentation that ensures all subtasks can be fully offloaded to the Multi-access Edge Computing (MEC) infrastructure before the vehicle exits the current coverage area. This approach mitigates the issue of task offloading failure due to vehicular mobility. Cao et al. [1] proposed a scheme to divide onboard tasks into independent subtasks, reducing the volume of data to be offloaded to the MEC. Ji and Jiang [19] considered the dependencies between subtasks and used genetic algorithms to solve the multi-site cooperative computational offloading problem.This minimizes the time overhead associated with the offloading of individual subtasks.Our work builds upon these foundations by introducing a novel speed-sensitive task offloading scheme designed specifically for fleets of vehicles. Unlike previous studies, our focus is on the prioritization and scheduling of multiple task offloading scenarios to ensure successful offloading before vehicles leave the coverage area. Moreover, our approach incorporates a real-time position updating mechanism that leverages MEC collaboration to enhance the success rate of task return, even as vehicles move across different coverage areas. Sun et al. [13] proposed predicting the position of a vehicle at the time of task completion based on factors such as running speed and trajectory, and introduced two distinct offloading strategies. Their work highlights the need for predictive algorithms to manage the offloading process effectively. Our research complements this by providing a solution that not only predicts but actively manages the offloading process in real time, taking into account the dynamic nature of vehicular networks. Ning et al. [9] proposes a traffic control system deployed at the base station. Utilizing the feature that the coverage of the base station is larger than the RSU, it obtains the real-time position information of the vehicle and returns the results of the task processing to the vehicle. Li et al. [7] also adopts a centralized control approach to determine whether the processing results are returned successfully or not based on the signal-to-noise ratio between the vehicle and the RSU. However, the centralized control approach not only introduces additional transmission delay, but also reduces the flexibility of system deployment and becomes another form of "cloud" [10]. Therefore, this paper proposes a dynamically changing multi-level MEC architecture, which increases the flexibility of system deployment while improving the success rate of backhauling through MEC collaboration instead of cloud-side collaboration.

In contrast to previous research, we focus on sequencing the fleet's tasks to ensure that offloading is completed in a timely manner before the vehicles leave the coverage area. This approach addresses the problem of failed task offloading due to vehicle movement. We introduce a collaborative backhaul system that leverages cooperation between MECs to improve the success rate of processing result returns. This addresses the problem of failed result returns when vehicles leave the RSU coverage area. We have improved the genetic algorithm specifically for task prioritisation and scheduling, which improves the responsiveness and flexibility of the system.

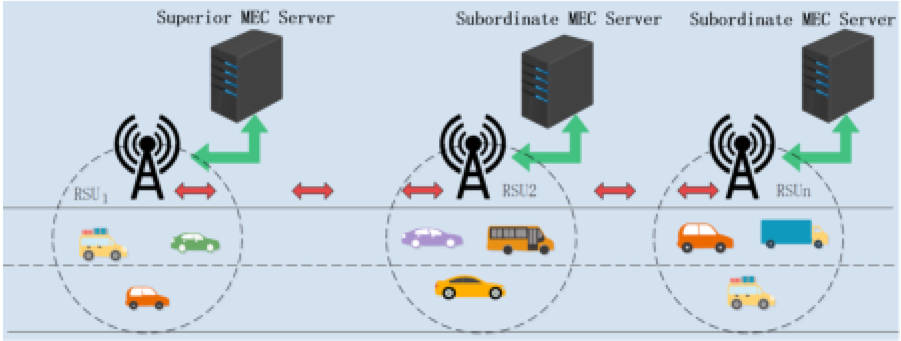


Fig. 3.1: Superior MEC and a subordinate MEC

Our solution ensures that vehicles can quickly offload tasks and receive results even in high-mobility situations, resulting in a safer and more responsive autonomous driving experience.

In conclusion, previous research has not fully addressed the operational challenges specific to in-vehicle networks, such as the need for real-time location updates, scalability, computational complexity, and security and privacy concerns. In this paper, we investigate speed-sensitive task offloading and co-return schemes in vehicular networks (VNETs) and introduce a dynamically evolving multi-layer MEC architecture that optimises the efficiency of task offloading and significantly improves the success rate of task return, thus providing a solution to the challenges posed by vehicular mobility in 5G and Internet of Things (IoT) environments.

3. System Architecture and Achemes.

3.1. Dynamically Changing Multi-level MEC Architecture. In vehicular networking, Road Side Units (RSUs) have a specific coverage area, and when a vehicle enters an RSU's coverage, it establishes a Vehicle-to-Infrastructure (V2I) connection. The Multi-access Edge Computing (MEC) server connected to the RSU provides computing resources to vehicles within the coverage area. As vehicles leave one RSU's coverage, they connect to the next RSU. RSUs also establish wired or wireless connections with other RSUs, creating a larger connection range than the V2I range. As shown in Figure 3.1, when a vehicle enters an RSU's coverage and generates a computational task, the MEC server for that RSU becomes the superior MEC, while other connected MEC servers become subordinate MECs. The superior MEC decides on task offloading, while subordinate MECs provide information on local link status, vehicle position, and computational capacity. Each MEC acts as both a superior and subordinate MEC, forming a dynamic hierarchical structure. This structure facilitates improved coordination and resource allocation. The superior MEC decides whether to offload a task to itself or to a subordinate MEC based on computational capacity, link quality, and vehicle velocity. It also coordinates with other superior MECs to ensure seamless task offloading and result return as vehicles move across different RSU coverage areas. The hierarchical structure optimizes task offloading and result return success rates, enabling informed decisions based on real-time information from subordinate MECs, adapting to vehicle mobility and optimizing resource utilization. Collaboration between superior and subordinate MECs ensures service continuity and minimizes average task offloading time by dynamically allocating tasks to the most suitable MEC based on network conditions and vehicle trajectory. Proactive task offloading by the superior MEC to subordinate MECs in anticipation of vehicle movement reduces the likelihood of offloading failure due to exiting RSU coverage. In summary, the dynamically changing multi-level MEC architecture provides a flexible and scalable framework for managing computational tasks in vehicular networks, effectively handling high mobility challenges and ensuring seamless service delivery.

3.2. Multi-level MEC architecture sorts offloading tasks. At moment t, there are multiple vehicles within the coverage area of a single RSU, and each vehicle may generate computational tasks. It is assumed that the communication mode between the RSU and the vehicles uses time-division multiplexing. Due to the limited link resources, the superior MEC must sort all the tasks generated at the moment, as shown in Fig. 3.2.

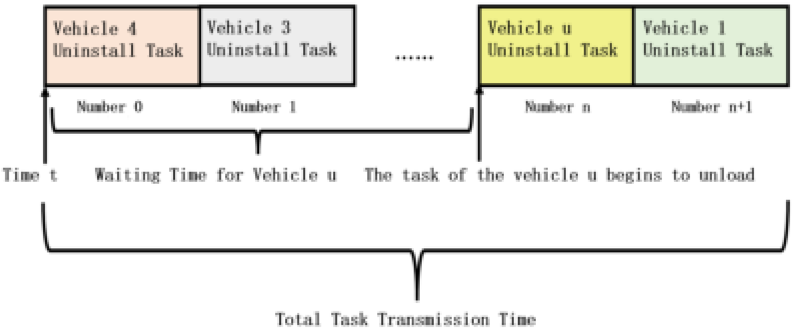


Fig. 3.2: Sorting of tasks by superior MECs

The task with the serial number n must wait for the previous n-1 task to finish offloading before it can be offloaded. Since the vehicle is always on the move, the position of the vehicle corresponding to the task number n has changed when the previous n-1 task has finished offloading, therefore, the following effects may be caused: 1) the distance between the vehicle and the RSU changes: the change in distance will lead to the change in the link state between the vehicle and the RSU, resulting in the change of the task n offloading time, and ultimately leading to the change of the average task offloading time; 2) the vehicle moves out of the current RSU coverage area, and the average task offloading time changes.

Therefore, the superior MEC needs to consider how to sort all the tasks generated at a moment in time t to reduce the average task offloading time while ensuring that the vehicle does not drive out of the current coverage area.

4. System Schemes. In the future intelligent transportation system (ITS), there will be a lot of negotiations between vehicles and vehicles (V2V) and vehicles and roads (V2I), which will generate a lot of task offloading and backhaul optimization problems. In the process of data transmission between vehicles, multilateral server vehicles, etc., many data services must be performed within a specified time, such as path planning services, information, entertainment and information transmission services. Therefore, the Internet of Vehicles needs to focus on how to reduce the latency of network data transmission. The use of edge computing resources, as well as the use of network function virtualization and software defense network technology, can greatly improve the efficiency and quality of end users to obtain services and reduce latency and energy consumption in data processing.

We define vehicles as autonomous agents that can act as temporary storage and computation resources for nearby vehicles, especially when the central server is overloaded or unavailable. Vehicles can perform temporary storage and computation tasks for other vehicles during peak traffic periods or in areas with limited MEC server coverage. Roadside Units (RSUs) act as communication nodes that collect information from the vehicles and forward it to the appropriate MEC servers, providing real-time updates about link status and computational capacity. As shown in Table 4.1, the differences between traditional and vehicular network operations are summarised.Strategically deployed Mobile Edge Computing (MEC) servers handle offloaded tasks and can operate in a centralised or decentralised manner, managing tasks over a large area or within a smaller geographic area, respectively. The task offloading mechanism utilises a genetic algorithm executed by the centralised MEC servers, which makes informed decisions on task offloading based on information provided by RSUs and vehicles.

4.1. Speed Sensitive Task Offloading Based Scheme. Speed Sensitive Computing Offloading Model (SSCOM), a speed-aware task offloading scheme based on the classical offloading model, adds the consideration of vehicle speed and the constraints of the vehicle crossing the zone situation. The offloading model is as shown in Fig. 4.1, this scheme uses the highway as the horizontal coordinate. the RSU is located on one side of the highway, and the perpendicular distance from the highway is y_0. Assuming that the coverage area of the

Table 4.1: Differences between traditional and in-vehicle network operations

Traditional Network Operations	Vehicular Network Operations
Centralized servers provide all services	Centralized servers, decentralized nodes, and autonomous agents
Static network topology	Highly dynamic and mobile network topology
Limited peer-to-peer communication	Extensive peer-to-peer communication and cooperation
Fixed resource allocation	Adaptive resource allocation based on vehicle density and mobility
Predominantly static security measures	Dynamic security and privacy measures that adapt to the changing network environment

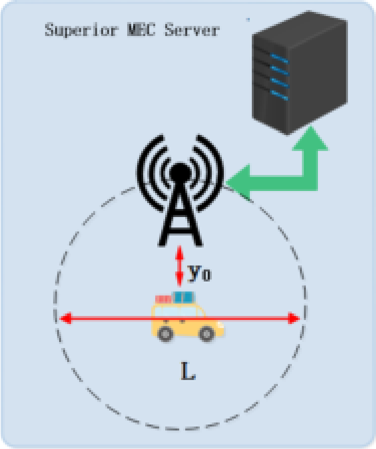


Fig. 4.1: Internet of Vehicles Scenario

RSU is L, the left boundary of the coverage area of the RSU is used as the starting point of the horizontal coordinate. Therefore, the coordinates of this RSU are $(\frac{L}{2}, y_0)$. Assume that all vehicles will not collide, the vertical coordinates are all 0, the speed of vehicle is v_u, and all vehicles are traveling at a uniform speed within the coverage area of the RSU.

At moment t, vehicle u generates a computational task that needs to be offloaded. At this time, the horizontal distance between vehicle u and the left boundary of the RSU is $X_u(t)$ and the vertical distance is y_0. At the same moment, there are other vehicles that generate tasks, and their task offloading order is prioritized over the vehicle u. The set of these vehicles is $pre(u)$. Therefore, the vehicle has to wait for a time of $\Delta_u = \sum_{m \in pre(u)} T_m$. The distance traveled by the vehicle during the waiting time is $v_u \Delta_u$. At this time, the vehicle's horizontal coordinate is $X_u(t + \Delta_u) = X_u(t) + v_u \Delta(u)$ and the distance from the RSU can be expressed as:

$$r_u(t) = \sqrt{y_0^2 + (x_u(t + \Delta_u) - \frac{L}{2})^2} \tag{4.1}$$

At time $t + \Delta_u$, the data transmission rate between the vehicle and the RSU is:

$$R_u(t) = B \log(1 + \frac{P_u h_u(t)}{\sigma_u^2}) \tag{4.2}$$

where B is the bandwidth of the link channel, P_u is the transmit power of vehicle u, σ_u^2 is the Gaussian white noise power, and h_u is the channel gain parameter. h_u is related to $r_u(t + \Delta_u)$:

$$h_u(t) = \frac{h^2}{r_u(t + \Delta_u)^\delta} \tag{4.3}$$

where h is the channel fading factor of the upload link and δ is the path loss factor. It is assumed that the position of vehicle u does not change during the offloading of its own task. Vehicle u consumes the time to offload the task of size D_u at moment $t+\Delta_u$ as:

$$T_u(t)=\frac{D_u}{R_u(t)}=\frac{D_u}{BN\log\left(1+\frac{P_u\frac{h^2}{\sqrt{y_0^2+(x_u(t)+v_u\sum_{m\in pre(u)}T_m-\frac{L}{2})}^{\sigma}}}{\sigma_u^2}\right)} \tag{4.4}$$

Assuming that the set of vehicles at time t is U and the number of tasks is N (each vehicle generates one task), the average task offloading time is:

$$\overline{T(t)}=\sum_u^U\frac{T_u(t)}{N}=\sum_u^U\frac{D_u}{BN\log\left(1+\frac{P_u\frac{h^2}{\sqrt{y_0^2+(x_u(t)+v_u\sum_{m\in pre(u)}T_m-\frac{L}{2})}^{\sigma}}}{\sigma_u^2}\right)} \tag{4.5}$$

Assuming that the superior MEC offloads the tasks using the sorting method A , the following optimization objective is in place:

$$\min_A\overline{T(t)} \tag{4.6}$$

In prevent the vehicle from driving out of the coverage area of the RSU at the moment , the superior MEC also needs to consider the following constraints during the sequencing:

$$X_u(t+\Delta_u)\leq L \tag{4.7}$$

Assuming that there are a total of N tasks at time t, there are $N!$ possible sorting methods. Therefore, in this paper, a genetic algorithm is used to solve the optimal offloading sorting scheme.

4.2. MEC Collaboration-Based Processing Result Return Scheme. During the time the MEC is processing the task, the vehicle may have already driven out of the RSU's coverage area. Even if the MEC is able to predict the vehicle's position from the vehicle's historical travel speed and return the result via inter-RSU communication, there is no guarantee that the vehicle can be found at the target MEC.

Therefore, based on the unpredictability of vehicle user behavior, this paper introduces a real-time position update mechanism on MECs. A mapping table is maintained on each MEC. The table records the task currently being processed, and the RSU serial number of the task corresponding to the latest passing of the vehicle.

Assume that the vehicle has successfully offloaded the task to the superior MEC. The vehicle remains on an unpredictable driving trajectory while the superior MEC is processing the in-vehicle task. Therefore, the superior MEC divides the processing time of the task into multiple fixed-size time slots τ. At each time slot τ, the superior MEC sends a vehicle position update request message to the subordinate MEC, and keeps updating the serial number of the RSU in which the vehicle is located. When the task processing is finished, the superior MEC transmits the processing result back to the vehicle via inter-RSU communication based on the updated position information. The MEC acting as the superior MEC is itself a subordinate MEC to other MECs, so this MEC also responds to vehicle position update requests from other superior MECs. The superior MEC processing is shown in Fig. 4.2.

Assume that the vehicle u offloads tasks Ku to the superior MEC. The superior MEC initiates position requests for the vehicle at fixed intervals. The data volume of the position request is very small, and thus imposes a negligible transmission delay. The superior MEC first checks whether the vehicle is in the coverage area of its own RSU. If it is, it updates the value in the mapping table; if it is not, it obtains the vehicle information from the subordinate MEC corresponding to the RSU through the corresponding RSU serial number in the mapping table. If the subordinate MEC also has no vehicle information, it indicates that the vehicle has driven out of the current range. However, since the subordinate MEC is the latest position of the vehicle during the last time. So it is easier to get the vehicle's position information by initiating a query from that lower.

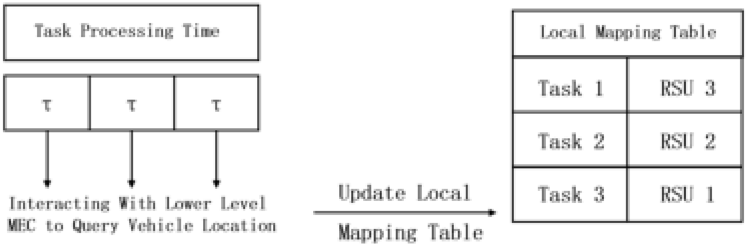


Fig. 4.2: Position of the vehicle corresponding to the superior MEC update task

4.3. Overcoming Algorithm Efficiency and Resource Management Challenges. The SSCOM solution addresses the challenges of algorithm efficiency and resource management in a dynamic vehicle environment. In order to reduce the computational intensity of the genetic algorithm used for task prioritisation and scheduling, parameters are optimised to achieve a balance between computational efficiency and solution quality. SSCOM dynamically assigns tasks to the most appropriate MEC servers based on network conditions and computational power, using a hierarchical MEC architecture where the upper-level MECs make informed decisions about task offloading, while lower-level MECs provide real-time information. This ensures efficient task execution and resource utilisation, minimising latency. The dynamically changing multi-level MEC architecture of the scheme facilitates seamless task offloading and result return, adapting to the high mobility and dynamic nature of in-vehicle networks. The genetic algorithm has a computational complexity of $O(10^8)$, which makes it feasible to deploy in real vehicle networks considering the modest memory and CPU requirements. Thus, SSCOM ensures efficient task offloading and resource utilisation, making it a practical solution for vehicular networks.

5. Experiments. Based on the above system scheme, this paper conducts simulation experiments on the offloading scheme. The optimization objective of the speed-aware task offloading scheme is to minimize the average task offloading time of the task under the constraint that all vehicles do not drive out of the RSU range. For the NP problem, this scheme is solved using a genetic algorithm. It is assumed that the number of genes N on the chromosome represents the number of momentary tasks. Each gene represents the vehicle number that generates the task. One arrangement of genes on the chromosome represents one possible arrangement of task offloading. The specific process of the experiment is as follows.

5.1. Parameter Selection.

5.1.1. Parameter Selection. The parameters of the genetic algorithm have a significant impact on the performance of the SSCOM scheme, particularly the mutation probability, replacement probability, and the number of genetic iterations. To achieve optimal performance, we employ the method of controlled variables to study the effects of these parameters on the average task offloading time.

a) Mutation Probability: The specific settings are the initial population size of 1000, the number of genes (vehicles generating tasks) is 100, the number of iterations is 1000, and the mutation probability varies over the range [0, 0.1]. As shown in Figure 5.1, the average task offloading time tends to decrease within the range [0, 0.04] and then increases with further increases in the mutation probability. Therefore, the mutation probability chosen for this scheme is 0.04 (Figure 5.1)

b) Replacement Probability: The specific settings are the initial population size of 1000, the number of genes is 100, the number of iterations is 1000, and the replacement probability varies over the range [0, 1]. As depicted in Figure 5.2, the replacement probability chosen for this scheme is 0.5, as it provides the optimal balance between introducing new chromosomes and maintaining diversity in the population.

c) Number of Genetic Iterations: The specific settings were the initial population size set to 1000, the maximum number of iterations set to 1500, the number of genes set to 40, 80, and 120, respectively,

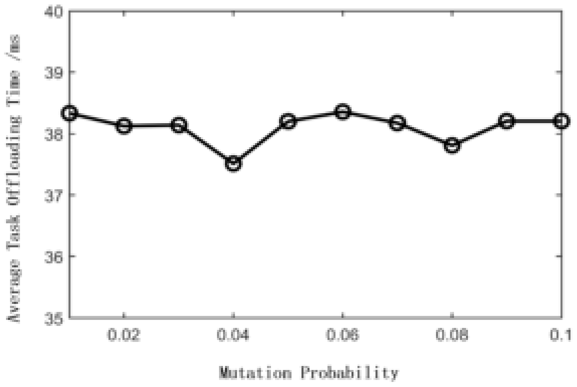


Fig. 5.1: Mutation probability distribution

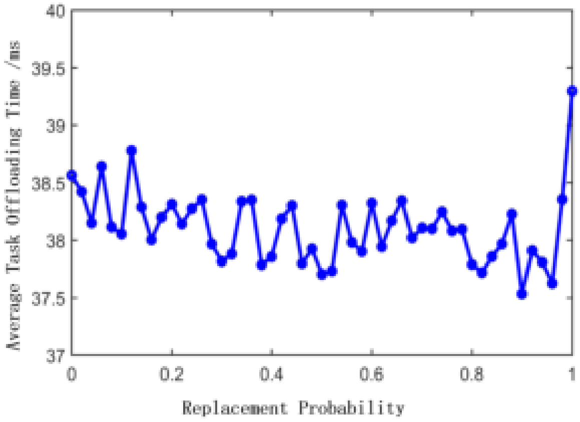


Fig. 5.2: Replacement probability distribution

the mutation probability is 0.04, and the replacement probability is 0.5. The number of iterations that yields the most stable performance without excessive computational overhead is selected. It is determined that the optimal number of iterations is 1000, balancing the trade-off between convergence speed and solution quality.

5.2. Test Process and Results.

5.2.1. Program Selection and Setup. Based on the parameter selection experiments described in Section 5.1.1, we now conduct the main simulation experiments to evaluate the SSCOM scheme. The experimental parameters involved in this experiment are shown in Table 5.1.

5.2.2. Algorithm Performance. To evaluate the performance of the SSCOM scheme, we vary the number of vehicles and observe the changes in the average task offloading time. As illustrated in Figure 5.3, when the number of vehicles is low, the difference between the schemes is minimal due to the limited number of tasks. However, as the number of vehicles increases, the SSCOM scheme achieves the optimal average task offloading time relative to the other schemes. For instance, at a high density of 120 vehicles, the average offloading time for the SSCOM scheme is 37 ms, which is 32% lower than the random offloading scheme (49 ms), 18% lower than the speed-first offloading scheme (45 ms), and 16% lower than the data-volume-first offloading scheme (43 ms).

Table 5.1: Parameters of simulation experiment

Experimental Parameters	Value
Link channel bandwidth B	10MHz
Vehicle transmit power P_u	500mW
Channel fading factor h	1
Path Loss Factor δ	4
RSU Coverage L	400m
RSU distance from y_0	10m
Gaussian white noise power σ_u^2	-100dB
Vehicle speed v_u	[16, 32]m/s
Number of vehicles N	[10, 150]
Task data size D_u	[0.1, 3.1]MB
Number of populations	1000
Maximum number of iterations	1000
Mutation probability	0.04
Replacement probability y_0	0.5

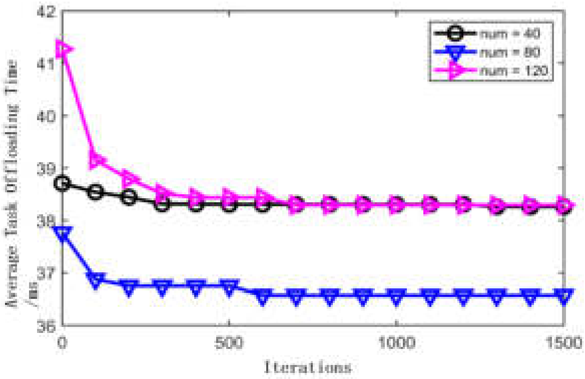


Fig. 5.3: Effect of the number of iterations on the performance of the scheme

5.2.3. Computational Resource Requirements. The SSCOM scheme requires computational resources for the genetic algorithm to run, which includes memory for storing the population and fitness values, and CPU cycles for performing the genetic operations. The memory requirement is proportional to the population size and the number of genes, and the CPU requirement is proportional to the computational complexity of the algorithm. Given the parameters used in the SSCOM scheme (Table 5.1), the memory requirement is moderate, as the population size is 1000 and the number of genes is 100. The CPU requirement is also reasonable, given that modern computing hardware can efficiently handle the computational complexity of $O(10^8)$.

5.2.4. Comparative Analysis. To further understand the performance of the SSCOM scheme, we analyze the average vehicle traveling speed and the average data volume of tasks for each group of tasks in the case of a group of tasks with 10 vehicles when the number of vehicles is 100. As shown in Figure 5.4, the speed-first offloading scheme prioritizes tasks associated with faster vehicles, whereas Figure 5.5 demonstrates that the data volume priority offloading scheme favors tasks with larger data volumes. In contrast, the SSCOM scheme takes into account both the speed and data volume to optimize performance.

5.2.5. Testing Results. We conducted experiments with varying vehicle densities to examine the SSCOM scheme's behavior under different traffic conditions. The density ranged from low (20 vehicles) to high (120 vehicles) in the RSU coverage area. As shown in Figure 5.6, the SSCOM scheme consistently performed well, maintaining a lower average task offloading time compared to alternative schemes, even as the number of vehicles

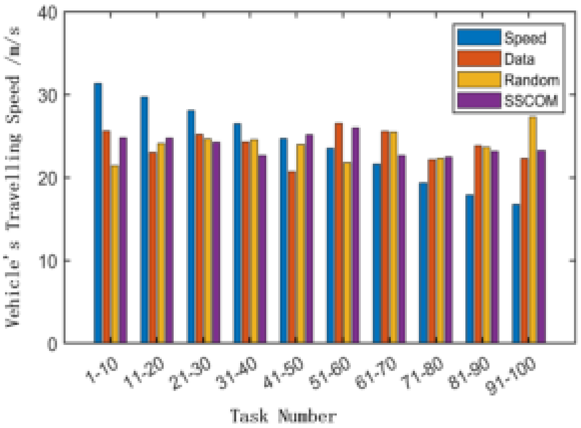


Fig. 5.4: The relationship between offloading scheme and vehicle speed

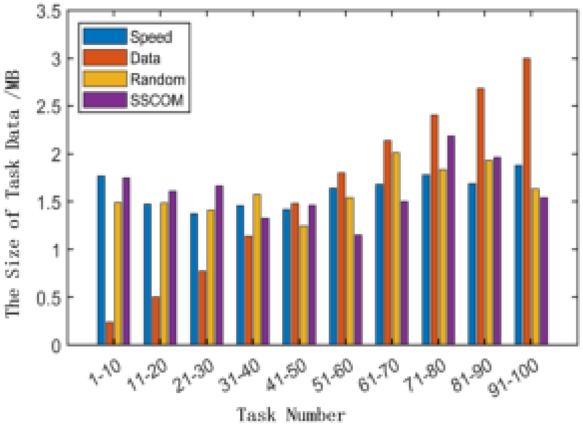


Fig. 5.5: The relationship between offloading scheme and task data volume

increased. Specifically, at a high density of 120 vehicles, the average offloading time for the SSCOM scheme was 37 ms, which is 32% lower than the random offloading scheme (49 ms), 18% lower than the speed-first offloading scheme (45 ms), and 16% lower than the data-volume-first offloading scheme (43 ms).

5.3. Analysis and Extension.

5.3.1. Detailed Analysis of the Genetic Algorithm. The genetic algorithm used in the SSCOM scheme is tailored to the specific needs of vehicular networks. The optimization objective is to minimize the average task offloading time of the task under the constraint that all vehicles do not drive out of the RSU range. As mentioned earlier, the number of genes on the chromosome represents the number of momentary tasks. Each gene represents the vehicle number that generates the task. One arrangement of genes on the chromosome represents one possible arrangement of task offloading. The genetic algorithm operates with a population size of 1000, a maximum number of iterations of 1000, and a mutation probability of 0.04. These parameters were selected based on the parameterization experiments described in Section 5.1.1.

5.3.2. Algorithm Complexity Analysis. The SSCOM scheme utilizes a genetic algorithm for task prioritization and scheduling. The complexity of the genetic algorithm is influenced by the number of genes, the population size, and the number of iterations. The computational complexity of the genetic algorithm can

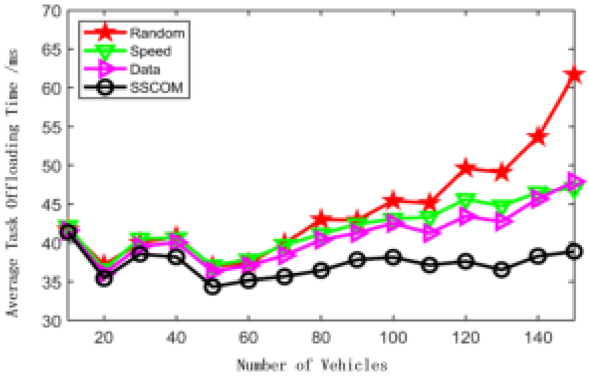


Fig. 5.6: Comparison of various offloading schemes

be estimated as $O(N * P * I)$, where N is the number of genes, P is the population size, and I is the number of iterations. For the SSCOM scheme, the number of genes (N) is set to 100, the population size (P) is set to 1000, and the number of iterations (I) is set to 1000. Therefore, the computational complexity of the genetic algorithm in the SSCOM scheme is approximately $O(10^8)$.

5.3.3. Scalability and Practicality. The SSCOM scheme is designed to be scalable and practical. The genetic algorithm used for task prioritization and scheduling has a computational complexity of $O(10^8)$, which is feasible for deployment in realistic vehicular networks. The memory and CPU requirements are moderate and can be met by current computing hardware. The SSCOM scheme ensures efficient task offloading and resource utilization, making it a practical solution for vehicular networks.

6. Operational issues. To address the operational complexities inherent in vehicular networks, we concentrate on five key business challenges: task offloading at varying speeds, collaborative backhaul during coverage range transitions, real-time location updates, scalability and computational complexity, and security and privacy. These challenges are critical to the effective operation of vehicular networks and require comprehensive solutions.

6.1. Security and Privacy. To ensure data security and privacy, we propose implementing centralized solutions that incorporate secure data transfer protocols, specifically Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS), to encrypt data both in transit and at rest. Robust access control mechanisms are essential to regulate who can offload tasks and receive results. Authentication mechanisms are included to verify the identities of vehicles and users, preventing unauthorized access. We also recommend encrypting and protecting both static and dynamic data using end-to-end encryption for data transmission and strong encryption algorithms, such as the Advanced Encryption Standard (AES) or Elliptic Curve Cryptography (ECC), for storing data. Mutual authentication protocols should be implemented between superior and subordinate MEC (Mobile Edge Computing) nodes to ensure that only legitimate MEC nodes can communicate, and that data is returned only to vehicle-related MEC nodes.

6.2. Realistic Validation. For realistic validation, we propose a comprehensive experimental setup that includes both simulations and field trials. Extensive simulations using real-world vehicular network scenarios will evaluate the performance of our proposed scheme. Field trials, conducted in collaboration with industry partners, will assess the practicality and effectiveness of the solution under realistic conditions.

6.3. Real-Time Location Updates. To integrate real-time location update algorithms into a multi-tier MEC architecture, we propose a scalable and flexible approach involving deploying real-time location services on MEC servers. These services update vehicle locations as they move within the network, ensuring that all relevant MEC servers are synchronized. This maintains the operational integrity of the vehicular network and supports the efficient management of tasks and resources.

7. Conclusion. This paper has presented a dynamically changing multi-level MEC architecture aimed at minimizing average task offloading times without compromising on-service continuity for vehicles traversing RSU coverage areas. By integrating vehicle speed and crossing scenarios into the classical offloading model, our proposed Speed Sensitive Collaborative Offloading and Backhaul (SSCOM) scheme has demonstrated a superior capability to minimize offloading times and enhance the task return success rate by at least 5%, outperforming conventional methods.It‘s practical significance lies in facilitating seamless and efficient service delivery in IoT environments, where timely data processing and reliable information feedback are critical. The SSCOM scheme ensures that vehicles can swiftly offload tasks and receive processed results even amidst high mobility, contributing to safer and more responsive autonomous driving experiences.

However, this study also has its limitations. Firstly, the parameter settings and scenario design in the simulation experiments are relatively simplified; future work should further validate the effectiveness of the proposed schemes in more complex real-world environments. Secondly, this paper’s schemes primarily focus on the performance of task offloading and result return, with less in-depth discussion on security and privacy concerns during the task offloading process. In future work, we intend to incorporate security mechanisms to ensure the safe transmission of data and the protection of privacy.Looking ahead, we anticipate that research in the field of vehicular networking will progress towards more intelligent, service-oriented, and secure development. With the proliferation of 5G technology and the enhancement of edge computing capabilities, vehicular networks will be able to offer richer and more efficient services.

8. Acknowledgements. My work was supported by the self-funded science and technology Innovation Project of Foshan City in 2022 (ID:2220001005675),Guangdong Provincial Department of Education Youth Innovation Talent Project (ID:2021KQNC-X123), Guangdong Software and Information Service Engineering Technology Research Center (ID:2022E07) and ”Climbing” Program (pdjh2022b0607) from Guangdong Province.

REFERENCES

[1] J. Cao, L. Yang, and J. Cao, *Revisiting computation partitioning in future 5g-based edge computing environments*, IEEE Internet of Things Journal, 6 (2019), pp. 2427–2438.

[2] C. Chen, C. Wang, T. Qiu, M. Atiquzzaman, and D. O. Wu, *Caching in vehicular named data networking: Architecture, schemes and future directions*, IEEE Communications Surveys and Tutorials, 22 (2020), pp. 2378–2407.

[3] L. Chen, S. Zhou, and J. Xu, *Computation peer offloading for energy-constrained mobile edge computing in small-cell networks*, IEEE/ACM Transactions on Networking, 26 (2018), pp. 1619–1632.

[4] R. Dhanare, K. K. Nagwanshi, S. Varma, and S. Pathak, *The future of internet of vehicle : Challenges and applications*, in 2021 International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 023–026.

[5] L. Han, M. Dong, and Song, *Service-based virtual network function placement algorithm in vehicle networking*, High-tech Communication, 31 (2021), pp. 341–349.

[6] X. Huang, L. He, X. Chen, G. Liu, and F. Li, *A more refined mobile edge cache replacement scheme for adaptive video streaming with mutual cooperation in multi-mec servers*, in 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1–6.

[7] M. Li, J. Gao, L. Zhao, and X. Shen, *Deep reinforcement learning for collaborative edge computing in vehicular networks*, IEEE Transactions on Cognitive Communications and Networking, 6 (2020), pp. 1122–1135.

[8] G. Liu, F. Dai, B. Huang, Z. Qiang, S. Wang, et al., *A collaborative computation and dependency-aware task offloading method for vehicular edge computing: a reinforcement learning approach*, Journal of Cloud Computing, 11 (2022), p. 68.

[9] Z. Ning, K. Zhang, X. Wang, M. S. Obaidat, L. Guo, et al., *Joint computing and caching in 5g-envisioned internet of vehicles: A deep reinforcement learning-based traffic control system*, IEEE Transactions on Intelligent Transportation Systems, 22 (2020), pp. 5201–5212.

[10] X. P.W.Ye and C. X.R.Yang, *Multi-agent reinforcement learning edge-cloud collaborative unloading for vehicle networking*, Computer Engineering, 47 (2021), pp. 13–20.

[11] S. Raza, W. Liu, M. Ahmed, M. R. Anwar, M. A. Mirza, et al., *An efficient task offloading scheme in vehicular edge computing*, Journal of Cloud Computing, 9 (2020), pp. 1–14.

[12] Z. Sheng, C. Mahapatra, V. C. M. Leung, M. Chen, and P. K. Sahu, *Energy efficient cooperative computing in mobile wireless sensor networks*, IEEE Transactions on Cloud Computing, 6 (2018), pp. 114–126.

[13] Y. Sun, X. Guo, J. Song, S. Zhou, Z. Jiang, X. Liu, and Z. Niu, *Adaptive learning-based task offloading for vehicular edge computing systems*, IEEE Transactions on Vehicular Technology, 68 (2019), pp. 3061–3074.

[14] S. Talal, W. S. M. Yousef, and B. Al-Fuhaidi, *Computation offloading algorithms in vehicular edge computing environment: A survey*, in 2021 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE), 2021, pp. 1–6.

[15] A. Talpur and M. Gurusamy, *Drld-sp: A deep-reinforcement-learning-based dynamic service placement in edge-enabled internet of vehicles*, IEEE Internet of Things Journal, 9 (2022), pp. 6239–6251.

[16] Z. Wu and D. Yan, *Deep reinforcement learning-based computation offloading for 5g vehicle-aware multi-access edge computing network*, China Communications, 18 (2021), pp. 26–41.

[17] C. You and K. Huang, *Exploiting non-causal cpu-state information for energy-efficient mobile cooperative computing*, IEEE Transactions on Wireless Communications, 17 (2018), pp. 4104–4117.

[18] H. Zhang, Z. Wang, and K. Liu, *V2x offloading and resource allocation in sdn-assisted mec-based vehicular networks*, China Communications, 17 (2020), pp. 266–283.

[19] Z.H.Ji and L.Y.Jiang, *A multi-site collaborative unloading algorithm based on genetic algorithm*, Computer Engineering and Science, 43 (2021), pp. 426–434.

[20] J. Zhou, D. Tian, Y. Wang, Z. Sheng, X. Duan, et al., *Reliability-optimal cooperative communication and computing in connected vehicle systems*, IEEE Transactions on Mobile Computing, 19 (2019), pp. 1216–1232.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Jun 28, 2024
Accepted: Aug 26, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1048–1056, DOI 10.12694/scpe.v26i3.4222

STUDY ON COST-SHARING METHOD OF POWER GRID ENGINEERING OPERATION AND MAINTENANCE BASED ON DEMATEL METHOD AND RANDOM FOREST [*]

HE PUYU[†] ZHAO MENGZHU[‡] WANG QIAN[§] YU GUANGXIU[¶] ZHAO KUIYUN[‖] AND WANG CHAO[**]

Abstract. The transmission and distribution (T&D) tariff reform raises higher requirements for the control of operation and maintenance (O&M) cost in power grid engineering. By constructing a model analyzing the O&M cost influencing factors, the paper uses an elastic regression analysis and verifies the key influencing factors on O&M cost, including 14 items such as population size, electricity consumption, gross regional product (GRP). And the subjective and objective weights of each influencing factor are analyzed using the DEMATEL method and random forest. Finally, by calculating the combined weight of factors affecting the power grid O&M cost of each local municipal company, the method of O&M cost-sharing is proposed, which can effectively improve the efficiency of operation and maintenance and make the cost-sharing more balanced and reasonable. Through the empirical analysis, the feasibility of this method is proved, which provides an effective reference for the cost-sharing and control of power grid O&M.

Key words: operation and maintenance (O&M); DEMATEL method; Random forest; combination weight; cost-sharing

1. Introduction. In recent years, artificial intelligence algorithms, including random forest algorithm, logistic regression, SVM, neural network, etc., have been widely used in all walks of life. How to make better use of artificial intelligence algorithms to do a good job in enterprise cost management has gradually attracted the attention of experts and scholars. In order to implement the relevant requirements for deepening the reform of the electric power system, improve the depth and breadth of the reform, and continuously improve the reasonableness of the T&D tariffs, the National Development and Reform Commission (NDRC) issued a "Provincial Grid Transmission and Distribution Tariff Pricing Measures" in January 2020, which stipulates that provincial grid T&D tariffs should be approved in accordance with the "approved costs with reasonable returns", and puts forward the requirement of imposing a rate cap on the O&M cost. There are many provinces and regions with large differences in topography, natural environment and resource endowment in China. In particular, some regions have long transmission distances for power grids, which makes power grid operation difficult and O&M cost high. However, under the premise of the national control on provincial O&M cost, the way to make reasonable O&M cost-sharing so as to maximize the cost-effectiveness not only meets the requirements of the control, but also plays a maximum role in the stable operation of grids and the sustainable development of the region, which becomes the focus direction of the study.

At present, there are many studies on O&M cost-sharing and O&M cost influencing factors. Among the studies on O&M cost-sharing, Reference [1] used Lasso regression algorithm for the influencing factors of O&M cost, and then allocates cost to each individual project based on key performance criteria. Reference [2] put forward the analytic hierarchy process (AHP) -entropy weight method based on distribution network equipment asset operation and maintenance cost optimization allocation mode. Reference [3] used elastic network algorithm to screen the influencing factors of operation and maintenance cost, and the efficiency standards of key factors are determined. Secondly, the CRITIC method is selected for weight side calculation, and the cost allocation coefficient of a single project is obtained. Reference [4] studied the direct costs, indirect costs and other related

[*] This work was supported by the State Grid Corporation of China Technology Project (SGSCJY00JJJS2310034).

[†] State Grid SiChuan Economic Research Institute, Chengdu 610041, China.
[‡] State Grid Sichuan Electric Power Company, Chengdu 6100411, China.
[§] State Grid Sichuan Electric Power Company, Chengdu 6100411, China.
[¶] State Grid Sichuan Electric Power Company, Chengdu 6100411, China.
[‖] State Grid Sichuan Electric Power Company, Chengdu 6100411, China.
[**] State Grid Sichuan Electric Power Company, Chengdu 6100411, China.

costs incurred during the life cycle of assets and equipment, optimized the cost composition during the life cycle of assets, and thus improved the cost collection and allocation method. Reference [5] combined with the application mechanism of power grid operating standard cost, determined the differentiated allocation algorithm of operating standard cost based on DEMATEL and combined weighting method. Taking substation maintenance service as an example, the differentiated allocation of operating standard cost was realized. And among the studies on O&M costs influencing factors, Reference [6] proposes to use Bayesian Model Averaging (BMA)-improved Grey Relation Analysis in terms of economic factors, equipment factors, environmental factors, and network structure for the O&M cost influencing factors and finally identifies nine key factors. Reference [7] studies and determines the direct and indirect O&M cost influencing factors on extra-high voltage equipment, and then quantifies the influencing factors using correlation analysis, percentage analysis and other methods. In addition, some scholars have calculated and predicted the whole life cycle cost of power grid by means of least square method and principal component analysis method [8], and proposed the use of cloud mass movement technology to deepen the role of physical data chain, collect technical and value information at each stage of the whole life cycle of equipment, and form a security, efficiency and cost optimization control strategy in line with the staged goals of enterprises [9].

The current studies on O&M cost-sharing and the analysis of influencing factors are relatively limited, and the cost-sharing studies focus on a single project itself or a single equipment without the perspective of actual O&M management on local municipal companies. On this basis, from the perspective of management, we establish an influencing factor analysis model on O&M cost, and uses elastic regression analysis to identify the key O&M cost influencing factors, and then calculate the subjective and objective weights of the influencing factors based on DEMATEL method and random forest, and finally calculate the combination weight, and put forward the cost-sharing method on O&M for local municipal companies.

2. Analysis of the power grid O&M cost influencing factors.

2.1. Influencing factor recognition. China has issued a “Measures for the Supervision of T&D Tariff” and qualified the T&D tariff cost components, including depreciation expense and O&M cost (i.e., operation, maintenance and repair costs). From the level of T&D pricing, grid O&M cost is mainly composed of material cost, repair cost, labor cost and other operation costs; from the level of cost types, there are mainly composed of operation cost, maintenance cost and repair cost; and from the level of basic cost components, they are mainly composed of labor cost, material cost, machinery cost as well as other costs. O&M cost of power grid equipment are influenced by many different factors. Brainstorm method is used to collect and summarize the most extensive factors affecting the cost of power grid operation and maintenance by inviting experts from all aspects of power grid operation and maintenance to discuss. Besides, considering the availability of data, the factors are sorted out and summarized to determine the influencing factors that need further analysis and research, as shown in Table 2.1. O&M cost influencing factors mainly include social, technical, environmental factors.

2.2. Influencing factor analysis model construction. In the last century, American experts such as Ehrlich first proposed IPAT model to study the impact of social and economic conditions on the environment, including social, economic, population, technology and other factors on harmful gas emissions and various energy consumption. However, the basic logic of this model is that the contribution of the influencing factors to the impact on environment is the same [10]. Obviously it is impossible to be the same in practice, so many scholars at home and abroad have conducted more in-depth studies on it, and then optimized model was proposed, namely the STIRPAT model [11],[12]. The premise of the establishment of the multiple line regression model is a roughly linear relationship between the independent variables and the dependent variables, but based on the analysis on the O&M cost of the various influencing factors, some of the influencing factors and the O&M cost do not show a conventional linear relationship, and there is an obvious multicollinearity problem among the factors. Considering that O&M cost is similar to the principle of energy consumption in the STIRPAT model, the grid O&M cost influencing factor analysis model is constructed with reference to the STIRPAT model as shown in the following equation:

$$I = aP1^{b}P2^{c}A1^{d}\cdots E4^{s}t \tag{2.1}$$

Table 2.1: Elastic regression analysis results.

No	1st level influencing factor	2nd level influencing factor
1	Social factor	Population size
2		Urbanization rate
3		GRP
4		Consumer price index
5		Power supply area
6		Electricity consumption
7	Technical factor	Substation capacity
8		Line length
9		Average equipment running time
10		Average annual outage time
11		Average annual outage number
12		Equipment failure rate
13		Comprehensive line loss rate
14		Substation capacitance to load ratio
15	Environmental factor	Average temperature
16		Topographic situation
17		Population density

Where I is grid O&M cost; a is the model coefficient; t is the random error term; b, c, d, ..., s represent the model elasticity coefficients for each influencing factor, respectively. If we put logarithms in the model, it can be converted into a line regression model, and then the regression analysis can be utilized for an in-depth study on the influencing factors.

2.3. Influencing factor analysis. Generally, linear regression models is applied to analyze the influencing factors and Ordinary Least Squares (OLS) is usually used to calculate the parameters of the regression model, but this cannot solve the multicollinearity problem among the influencing factors. Elastic Net Regression (ENR) can both reduce the coefficients as in Ridge Regression and select features as in Lasso Regression [13]. The advantages of ENR over Ridge and Lasso Regressions are that ENR can automatically select features when the data is highly correlated, reduces the effect of matrix singularity, and can handle high-dimensional small sample data [14], [15]. Therefore, ENR is used to solve the multicollinearity among various influencing factors, and modeling analysis is carried out to determine the key influencing factors. The expression of the elastic network regression model is as follows:

$$\widehat{\beta}(Elasticnet) = \arg\min_{\beta}\left\{\|Y - X\beta\|^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}{\beta_j}^2\right\} \tag{2.2}$$

If set $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}, \lambda = \lambda_1 + \lambda_2$, then

$$\widehat{\beta}(Elasticnet) = \arg\min_{\beta}\left\{\|Y - X\beta\|^2 + \lambda\left[\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}{\beta_j}^2\right]\right\} \tag{2.3}$$

where $\alpha\sum\limits_{j=1}^{p}|\beta_j| + (1-\alpha)\sum\limits_{j=1}^{p}{\beta_j}^2$ is called the penalty term of the Elastic net.

3. Grid O&M cost-sharing model.

3.1. Influencing factor subjective weights determination. Graph theory and matrices are used in DEMATEL method, whose basic principle is to identify the logical relationships between the influencing factors in the system, and then calculate the degrees of influence and the degrees of being influenced between them

[16], [17], [18]. The method fully considers the impact of the influencing factors, and a more comprehensive and in-depth determination of the relationship between them. It can also evaluate and analyze the interrelationships between different factors and the extent to which they affect the outcome of the decision. Compared with other methods, DEMATEL method has higher accuracy and reliability. Therefore we use it to get the subjective weights between the influencing factors. The specific steps are as follows:

(1) The O&M cost influencing factors that need to be analyzed are identified firstly. Based on the results of the influencing factor analysis in Section 2, the influencing factors are expressed as followed:

$$F_1, F_2, \cdots, F_n \tag{3.1}$$

(2) A five-level scale method is used to represent the relationships between each influencing factor and different values are assigned. When there is no influence between two influencing factors, the value is 0; when weak influence, the value is 1; when medium influence, the value is 2; when strong influence, the value is 3, and when very strong influence, the value is 4. Thus we can determine the relationships between the influencing factors and O&M costs.

(3) The influence degrees between the influencing factors are represented by constructing matrices, and an n-order matrix is used as the direct influence matrix $C = (a_{ij})_{n\times n}$.

(4) Then we regularize the direct influence matrix N. N = S × C. The calculation formula of S in the above equation is

$$S = \frac{1}{\max\limits_{1\le t\le n} \sum\limits_{j=1}^{n} A_{ij}} \tag{3.2}$$

(5) The comprehensive influence matrix T is calculated.

(6) Finally we calculate the influence degree and the degree of being influenced by the O&M cost influencing factors, from which the weights are then calculated.

3.2. Influencing factor objective weights determination. In order to make the weights calculated in a more reasonable way, Random Forest Intelligent Algorithm is introduced, which is an integrated learning algorithm based on decision tree as a base learner [19], [20], [21]. The method is used for the importance assessment to determine the objective weights. The random forest model is constructed as follows:

(1) The first training set is obtained by taking M1 samples from the training sample set of the feature metrics by put-back sampling method. And the number of samples in the training sample set of the feature indicator is M, M1<M.

(2) D1 feature indicators are randomly selected from the feature indicators and a feature indicator set is formed. Then the feature indicator with the optimal classification ability is selected from the feature indicator set and split.

(3) A training set and feature metrics are used to generate a decision tree.

(4) Repeat the above steps I times to obtain a random forest model consisting of I decision trees.

(5) The importance of the feature indicators is assessed through the Gini index. The main process of assessment is as follows:

1) The number of feature indicators in the training sample is set as n; the number of decision trees is I, and the number of categories of evaluation results is C.

2) The Gini index of node q of the ith decision tree is calculated.

$$Gini_q^{(i)} = \sum_{C=1}^{C} \left(p_{qc}^{(i)}\right)^2 \tag{3.3}$$

where $p_{qc}^{(i)}$ is the proportion of categories c of evaluation results in node q.

3) The importance $VIM_{jq}^{(i)}$ of the feature indicator at node q of the ith decision tree is calculated.

$$VIM_{jq}^{(i)} = Gini_q^{(i)} - Gini_l^{(i)} - Gini_r^{(i)} \tag{3.4}$$

where $Gini_l^{(i)}$ and $Gini_r^{(i)}$ are the Gini indices of the two new nodes after the branching of node q, respectively.

4) If the nodes in which the feature matrix X_j in the ith decision tree are set Q, then the importance of the feature matrix in the ith decision tree is calculated as $VIM_j^{(i)} = \sum\limits_{q \in Q} VIM_{jq}^{(i)}$.
5) The total importance of the feature matrix in the I decision tree is calculated using $VIM_j = \sum\limits_{i=1}^{I} VIM_{jq}^{(i)}$.
6) The importance of the feature indicator X_j is normalized and the weight $VIM_{j,1}$ of the feature indicator X_j is calculated.

3.3. Sharing method. We use the DEMATEL method and random forest and analyze the subjective and objective weights of each influencing factor. In order to calculate the combination weight of each factor, the objective function is established and derived the optimal combination weight based on the principle of minimum identification information [22], and then according to the indicator value of each city, the comprehensive score of each city is calculated, and the proportion of each city is obtained after normalization and then the cost-sharing on O&M is identified. The O&M cost is calculated as follows:

$$\xi_k = \frac{a_i^k \omega_i}{\sum\limits_{k=1}^{r} a_i^k \omega_i} \tag{3.5}$$

Where ω_i^1 is subjective weight; ω_i^2 is object weight; ω_i is the optimal combination weight.

$$\omega_i = \frac{\left(\omega_i^1 \omega_i^2\right)^{1/2}}{\sum\limits_{i=1}^{n} \left(\omega_i^1 \omega_i^2\right)^{1/2}} \tag{3.6}$$

4. Empirical analysis.

4.1. Influencing factor analysis results. An empirical analysis is carried out with the O&M data of 18 municipalities in a province, and the data are collected through the statistical yearbook as well as the actual situation of on-site O&M. The dataset includes data on population size, electricity consumption, urbanization rate, GRP, consumer price index, substation capacity, line length, average equipment running time, average annual outage time and number, equipment failure rate, comprehensive line loss rate, substation capacitance load ratio, power supply area, average temperature, topographic situation, and population density from 2016-2022. Using the influencing factor model, the correlation of each influencing factor is first analyzed.

The correlation analysis of each influencing factor is carried out by using Pearson correlation coefficient, and the analysis results show that there is a high correlation. If we use the ordinary least squares method to calculate the model parameters, it is certain that there is the problem of multiple covariance. Then ENR is used and three influencing factors are excluded, including urbanization rate, average equipment running time, and average temperature, and fourteen items are retained. The main reason is that the results of ENR show that the model coefficients of urbanization rate, average equipment running time and average temperature are 0, as shown in Table 4.1. From the model coefficients, the influencing factors that have a greater impact on the O&M cost include electricity consumption, population density, line length, power supply area, and population size. The less influential factors are equipment failure rate, comprehensive line loss rate, and so on.

4.2. Combination weight calculation. Firstly, based on the data of 14 indexes of population size, electricity consumption, GRP, consumer price index, substation capacity, line length, average annual outage time as well as the number of outages, equipment failure rate, comprehensive line loss rate, substation capacitance to load ratio, power supply area, topographic situation, population density, and so on, we use the five-level scale method, and invite 10 experts in O&M to determine the relationship between the influencing factors. After several rounds of discussion, we get the direct influence matrix. According to the steps of the DEMATEL method described above, the degree of influence, the degree of being influenced, and the degree of centrality of

Table 4.1: Elastic regression analysis results.

No	2nd level influencing factor	Symbol	Model coefficients
1	Population size	P1	0.1972
2	Electricity consumption	P2	0.6344
3	Urbanization rate	A1	0.0000
4	GRP	A2	0.0111
5	Consumer price index	A3	-0.0908
6	Substation capacity	T1	0.1441
7	Line length	T2	0.2125
8	Average equipment running time	T3	0.0000
9	Average annual outage time	T4	-0.0103
10	Average annual number of power outages	T5	0.0105
11	Equipment failure rate	T6	-0.0100
12	Comprehensive line loss rate	T7	-0.0050
13	Substation capacitance to load ratio	T8	-0.0314
14	Power supply area	E1	-0.2044
15	Average temperature	E2	0.0000
16	Topographic situation	E3	-0.0180
17	Population density	E4	-0.2149

Table 4.2: Centrality and weight of influencing factors.

No	2nd level influencing factor	Centrality	Weight
1	Population size	1.7356	0.1300
2	Electricity consumption	1.387	0.1039
3	GRP	0.8937	0.0669
4	Consumer price index	0.4579	0.0343
5	Substation capacity	1.7102	0.1281
6	Line length	1.7525	0.1313
7	Average annual outage time	0.6362	0.0476
8	Average annual outages number	0.7892	0.0591
9	Equipment failure rate	0.7253	0.0543
10	Comprehensive line loss rate	0.3561	0.0267
11	Substation capacitance to load ratio	0.5156	0.0386
12	Power supply area	0.9717	0.0728
13	Topographic situation	0.5285	0.0396
14	Population density	0.8925	0.0668

each influencing factor can be calculated. The weight of each influencing factor of O&M cost can be calculated by normalizing the centrality degree as shown in Table 4.2.

The centrality of the influencing factor can reflect the status and importance of the factor in all the influencing factors. Generally, the greater the centrality, the more important the factor. From the analysis of the centrality and weights, it can be seen that the main factors affecting the O&M cost are population size, electricity consumption, substation capacity, line length and power supply area. Then the data of 14 factors such as population size, electricity consumption, GRP is used to construct the random forest regression model and calculate the importance degree. Since the parameter of the number of decision trees in the random forest has the greatest influence on the effect of the model, the optimal number of decision trees is obtained using the grid search method. The regression analysis is carried out under the optimal parameters, and the results of the importance analysis of the influencing factors are obtained. From the results of the importance analysis, it can be seen that electricity consumption, GRP, substation capacity, line length is more important influencing

Table 4.3: Subjective, objective weight and of combination weight.

No	2nd level influencing factor	Subjective weight	Objective weight	Combination weight
1	Population size	0.13	0.0310	0.078
2	Electricity consumption	0.1039	0.2295	0.189
3	GRP	0.0669	0.1438	0.12
4	Consumer price index	0.0343	0.0008	0.007
5	Substation capacity	0.1281	0.3229	0.249
6	Line length	0.1313	0.2087	0.203
7	Average annual outage time	0.0476	0.0005	0.006
8	Average annual outages number	0.0591	0.0001	0.003
9	Equipment failure rate	0.0543	0.0131	0.033
10	Comprehensive line loss rate	0.0267	0.0065	0.016
11	Substation capacitance to load ratio	0.0386	0.0039	0.015
12	Power supply area	0.0728	0.0010	0.011
13	Topographic situation	0.0396	0.0014	0.009
14	Population density	0.0668	0.0367	0.061

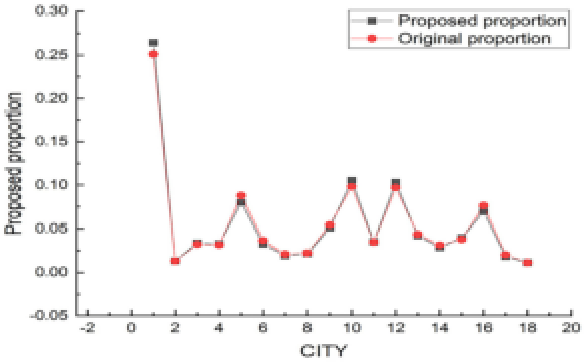


Fig. 4.1: Cost-sharing result

factors, and the consumer price index, the average annual power outage time and number are less important influencing factors.

4.3. Cost-sharing results. According to the subjective and objective weights, the combination weights of each O&M cost influencing factors are calculated as shown in Table 4.3. Based on the indicator values of each municipality, the combined score of each municipality is calculated, and the proportion of each municipality is derived after normalization. Finally the O&M cost is apportioned. The original proportion of cost-sharing and the proportion of this study are shown in Fig 4.1, the costs of city 1, city 10, city 12 should be increased, and the costs of city 5, city 6, city 9, city 14, city 16 and so on should be decreased. Through the further adjustment on O&M cost, it is possible to effectively improve the efficiency of O&M, and make the cost distribution more balanced and reasonable.

5. Conclusion. This paper analyzes the O&M cost influencing factors from three aspects: social factors, technical factors, and environmental factors. And the key factors are identified by constructing the STIRPAT influencing factor model. The key factors affecting O&M cost were determined by elastic regression analysis. By using the DEMATEL method and Random Forest, we put forward the O&M cost-sharing method for local and municipal companies. Finally using empirical analysis, we prove the feasibility of the method. The main results of this study include:

(1) With reference to the STIRPAT model, a model is constructed for analyzing the grid O&M influencing factors, and the problem of multiple covariance between the influencing factors is eliminated using elastic regression method. Through the analysis, the key factors are identified, of which the more influential influencing factors include electricity consumption, population density, line length, power supply area, population size, and substation capacity.

(2) Subjective weights are calculated using DEMATEL method, and objective weights are calculated using random forest, and the combination weights of the final O&M cost influencing factors are calculated, which shows that the factors that have a greater impact on the O&M cost include electricity consumption, GRP, substation capacity, and line length, and the factors that have a lesser impact include the annual consumer price index, topographic situation, the average annual power outage time and number.

(3) Combining the index value of each influencing factor of each city with the combination weights, the comprehensive score of each city can be derived. And after normalization, the proportion of O&M cost shared by each city can be derived, which provides a reference and basis for sharing and controlling the O&M cost of local municipal companies.

Due to the limitation of sample data, the analysis of this study is only limited to local municipal companies, and no comparative analysis of multiple provinces has been carried out. Further research is needed on whether the method is suitable for the cost management of provincial power grid companies.

The results of this study can guide the management of O&M cost of local municipal companies, and help to improve the economic benefits and technical level of operation and maintenance. Through further research and improvement, it can also be applied to the cost management of provincial power grid companies.

REFERENCES

[1] Wang Yongli, Wang Xiaohai, Wang Shuo, et al., *A Method Allocating Operation and Maintenance Cost of Power Grid Project Based on Transmission and Distribution Price Reform., Power System Technology, 44 (01), 332-339, 2020.*

[2] Tang Xuejun, Tan Zhongfu, Li Zhiwei, et al., *Distribution Network Equipment asset transportation and inspection cost optimization Model based on AHP-entropy weight Method., China Electric Power Construction, 43(10), 166-172, 2022.*

[3] Wang Yongli, Wang Shuo, Zheng Yan, et al., *Calculation of power grid operation and maintenance cost allocation based on elastic network. , Automation of Electric Power Systems, 44(20), 165-172, 2020.*

[4] Shen Wang., *Research on Asset Life Cycle Cost Collection and Allocation Method based on Power grid equipment. , Southern Energy Construction, 8 (S1), 53-58, 2021.*

[5] Wang Hongjin, Yu Zebang, Ren Yan, et al., *A differentiated cost allocation algorithm for power grid operating standards based on DEMATEL and Combinatorial weighting. , Electric Power Construction, 42 (08), 127-134, 2021.*

[6] Luo Chaoyueling, Li Zhiwei, Xu Zhenyu, et al., *Evaluation and Analysis of Factors Influencing the Operation and Inspection Costs of Distribution Network Equipment Assets. , Electric Power, 56 (07): 216-227, 2023.*

[7] Zhou Hongyu, Xue You, Liu Joyu, et al., *Measurement and Analysis of Impacting Factors for Operation and Maintenance Costs in UHV Substations. , Electric Power Construction, 2018, 39 (01): 19-29.*

[8] Xiong Zhiwei, Xiong Yuanxin, Xiong Yi., *Life cycle cost prediction of substation based on QPPO optimized LS-SVM. , Electrical Measurement & Instrumentation, 58 (06), 76-81, 2021.*

[9] Liu Dan, Liang Yiming, Lin Chuqiao, et al., *Research on Life Cycle Operation Information Collection and Visualization of Power Grid Main Equipment Assets. , Jilin Electric Power, 52 (02), 43-45, 2024.*

[10] Aziz, Ghazala, Sarwar, Suleman, Hussan, Muhammad Wasim, et al., *The importance of extended-STIRPAT in responding to the environmental footprint: Inclusion of environmental technologies and environmental taxation., Energy Strategy Reviews, 50, 2023.*

[11] Somoye, Oluwatoyin Abidemi, Ozdeser, Huseyin, et al., *The determinants of CO2 emissions in Brazil: The application of the STIRPAT model., Energy Sources, 45(4), 10843-10854, 2023.*

[12] Lund, Ibrar H., Shaikh, Faheemullah, Harijan, Khanji, et al., *Prospects of natural gas consumption in Pakistan: based on the LMDI-STIRPAT PLSR framework., Environmental Science and Pollution Research, 31(2), 2090-2103, 2024.*

[13] Alhamzawi, Rahim, Ali, Haithem Taha Mohammad., *The Bayesian elastic net regression., Communications in Statistics: Simulation and Computation, 47(4), 1168-1178, 2018.*

[14] Al-Jawarneh, Abdullah S., Ismail, Mohd Tahir, Awajan, et al., *Improving accuracy models using elastic net regression approach based on empirical mode decomposition., Communications in Statistics: Simulation and Computation, 51(7), 4006-4025, 2022.*

[15] Sloboda, Brian W., Pearson, Dennis, Etherton, Madi., *An application of the LASSO and elastic net regression to assess poverty and economic freedom on ECOWAS countries, Mathematical Biosciences and Engineering, 20(7), 12154-12168, 2023.*

[16] Sathyan, Rinu, Parthiban, P., Dhanalakshmi, R., Sachin, M.S., *An integrated Fuzzy MCDM approach for modelling and prioritising the enablers of responsiveness in automotive supply chain using Fuzzy DEMATEL, Fuzzy AHP and Fuzzy TOPSIS, Soft Computing, 27(1), 257-277, 2023.*

[17] Büyüközkan, Gülçin, Karabulut, Yağmur, Göçer, Fethullah., *Spherical fuzzy sets based integrated DEMATEL, ANP, VIKOR approach and its application for renewable energy selection in Turkey, Applied Soft Computing, 158, 2024.*

[18] Quezada, Luis E., López-Ospina, Héctor A., et al., *A Method for Formulating a Manufacturing Strategy Using Fuzzy DEMATEL and Fuzzy VIKOR., Engineering Management Journal, 36(2), 147-163, 2024.*

[19] Cadenas, Jose M., Garrido, M. Carmen, Martínez, Raquel, Bonissone, Piero P., *Extending information processing in a Fuzzy Random Forest ensemble. ,Soft Computing, 16(5), 845-861, 2012.*

[20] Levantesi, Susanna, Nigri, Andrea., *A random forest algorithm to improve the Lee-Carter mortality forecasting: impact on q-forward., Soft Computing, 24(12), 8553-8567, 2020.*

[21] Yoo, Byoung Hyun, Kim, Kwang Soo, Park, Jin Yu, et al., *Spatial portability of random forest models to estimate site-specific air temperature for prediction of emergence dates of the Asian Corn Borer in North Korea., Computers and Electronics in Agriculture, 199, 2022.*

[22] Wang Shi, Xu Lei, Ke Yuxian, Hu Kaijian., *Scheme optimization of supporting in deep underground roadway based on GRA-TOPSIS with optimal combination weight., Journal of Chongqing University, 42 (06): 78-87, 2019.*

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Jun 30, 2024
Accepted: Nov 29, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1057–1070, DOI 10.12694/scpe.v26i3.4225

SMART LAVATORY SOLUTION: INTEGRATING IOT AND DEEP LEARNING MODELS FOR ENHANCED HYGIENE

JIGNA PATEL[*], AESHWI SHAH[†], CHAUDHARI RUSHALI[‡], JITALI PATEL[§], AND VIJAY UKANI[¶]

Abstract. In the current era of smart technology, integrating the Internet of Things (IoT) with Artificial Intelligence has revolutionized several fields, including public health and sanitation. The smart lavatory solution proposed in this paper improves hygiene standards using deep learning models and IoT system. The proposed system collect real-time data from deployed sensors to monitor and assess hygiene conditions regularly. Proposed model consists of four consequent phases as hardware implementation, data preprocessing, application and user interface modules. Rasberry Pi based sensor integration at hardware layer, normalization based techniques at data preprocessing layer, LSTM and GRU based deep learning model at application development layer and mobile notification to the cleaning staff at user interface layer ensure efficiently cleaning and monitoring of lavatory systems. Prior to assessing the proposed model's testing accuracy experiments on the activation functions, optimizer, learning rate, and number of epochs were selected to choose the best to prevent overfitting or underfitting problems. With an accuracy of 98.61%, the proposed system performs better than the conventional approaches.

Key words: Internet of Things, Deep Learning, Long-Short Term Memory, Hygiene, Lavatory

1. Introduction. Traditional lavatory management system is a time consuming process. In order to resolve this issue, an IOT-based approach is introduced which performs real-time monitoring for automated cleaning schedules based on the requirements [22]. Additionally, the earlier system was time-consuming. To resolve these challenges, a smart lavatory is presented as a better solution. IOT-based smart lavatory extends the overall restroom experience in public spaces and proposes real-time monitoring for automated cleaning schedules based on the requirements. It also provides touchless features for improved hygiene, energy-efficient lighting control, occupancy indicators, enhanced user experiences with personalized settings, and integrated air quality sensors [21]. Also, these systems are suitable for high-traffic public areas like malls, airports, and offices since they optimize water and energy use while promoting sustainability. It can monitor washroom occupancy using the Passive Infrared sensor, track soap supplies, clean liquids, and toilet paper, monitor cleanliness levels using different IoT sensors, and send alerts if cleanliness or any other service is required [19, 5]. Moreover, this system enables the management to assign a worker to clean a specific area and remotely operate the cleaning system [18, 20].

The percentage of nations across each continent that have installed smart systems is depicted in Figure 1.1. Asia has the highest rate of acceptance (36.4%), followed by Europe (27.3%). Lower implementation rates are found in North America, Africa, and South America. As many Asian nations appear to be allowing the implementation of smart restroom technology, it indicates their success.

The common perception might be that most restroom germs and bacteria are found on the toilet seat but in reality, bacteria levels are much higher on the floor and high-touch surfaces, including the sink and faucet handles, hand dryers, light switches, doorknobs, as shown in the Figure 1.2. This in turn will spread a variety

[*] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India, (jignas.patel@nirmauni.ac.in)

[†] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (21bce012@nirmauni.ac.in)

[‡] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (21bce037@nirmauni.ac.in)

[§] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (jitali.patel@nirmauni.ac.in), Corresponding Author

[¶] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (vijay.ukani@nirmauni.ac.in)

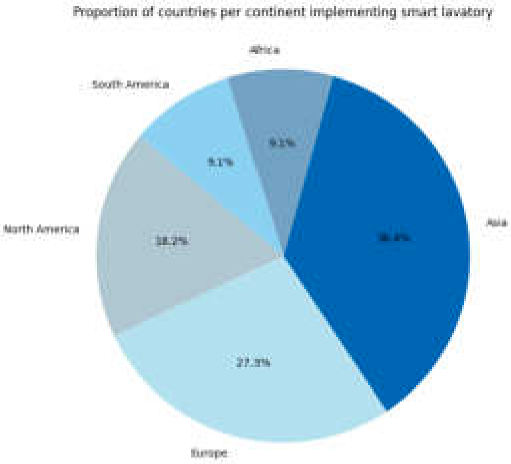


Fig. 1.1: Continents adopting smart lavatory systems [25]

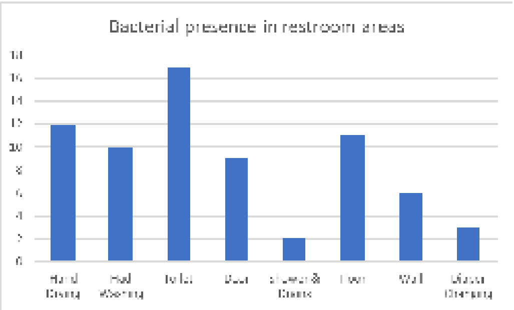


Fig. 1.2: Bacterial presence in restroom areas [27]

of diseases. It has been shown that hand dryers that blow warm air might be absorbing bacteria from the air, and dumping them on the newly washed hands of users. So Paper towels are the most hygienic way to dry hands. Studies have found that frequently touched surfaces such as soap dispensers and toilet handles have bacteria that are both skin-associated and fecal-originated, indicating that surface contact in restrooms is another important mechanism for transmission of illness [27].

Poor hygiene conditions are one of the main issues while using public toilets, as Figure 1.3 shows. The main issue, faced by 55% of users, is unclean restrooms. Other issues can be not having soap or toilet paper, lack of privacy, the toilets being occupied most of the time, etc. These problems show the urgent need to improve the ease of use and hygiene of public restrooms to improve user satisfaction and health [9].

The primary motivation for developing smart lavatory systems is the significant improvement in hygiene standards and user satisfaction [21, 17]. Due to the germs and bacteria found in public restrooms, people are prone to be affected with severe gastrointestinal distress, fever, and fatigue [26]. Table 1.1 shows that diarrheal diseases have a major national and international impact. These diseases appear to be the cause of 1.6 billion deaths worldwide, showing their broad effect on human health.

Users in public areas often face long wait times when using traditional restrooms. This is due to several factors, like inefficient usage, and a lack of real-time information about restroom availability. A key advantage of smart lavatory management is to minimize this queuing frustration by displaying real-time occupancy using the Passive Infrared sensor, and allowing users to report any issues using a real-time feedback system [18]. Additionally, data-driven cleaning schedules maximize the productivity of the cleaning crew by ensuring that

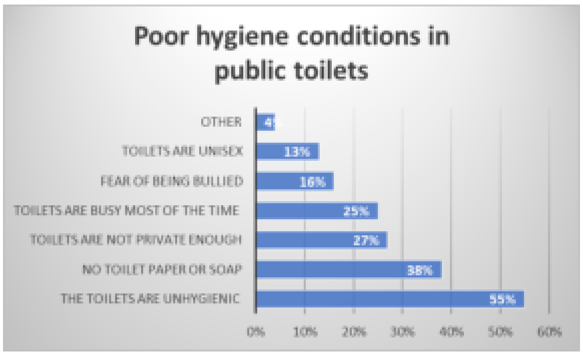


Fig. 1.3: Poor hygiene conditions in public toilets [9]

,

Table 1.1: Diseases related to poor hygiene[14]

Diseases	**Deaths(World)**	**Deaths(India)**
E. coli	2.8 million	1.6 million
Salmonella	1.5 million	0.68 million
noroviruses	2.1 million	19
Hepatitis A	3930	424
Diarrhoeal diseases	1.6 billion	2.2 million

bathrooms are cleaned only when necessary [18]. Water-efficient toilets are another crucial component of the system; they are designed with faucets that use less water for hand washing or flushing and reduce water waste by ensuring fixtures are only turned on when necessary [2]. Moreover, when the restroom is unoccupied, the motion sensors can automatically adjust the ventilation and lighting, saving energy [16].

This paper presents the implementation of a deep learning model, namely Long-Short Term Memory (LSTM), in conjunction with the Internet of Things (IoT) to create a smart lavatory management system. The system can alert janitors to clean the restrooms when needed, based on parameters such as gases, temperature, humidity, and occupancy, and detect foul smells. This approach enhances efficiency by optimizing cleaning schedules, saving both time and costs, while also promoting energy efficiency. We implemented this proposed work into practice in the specific washrooms of Nirma University's educational campus. The authors have ensured that notifications are provided correctly when needed and have reduced false alarms by using LSTM and GRU. Our smart restroom system's implementation of LSTM enhances its capacity to provide precise forecasts, guaranteeing better hygiene and upkeep while optimizing resource usage.

2. Related Work. The development of smart toilet technologies has been explored over the past years, as described in Table 2.1, by integrating IoT, machine learning, and deep learning to improve hygiene, user experience and resource efficiency. Lokman et. al. proposed an IoT-based smart toilet system that included genetic algorithms, ARIMA, KNN, and SVM models [11]. This helped to improve reliability and reduce cleaning and energy costs. However, this proposed model had limitations. Similarly, [7] aimed to elevate public health and sanitation by using micro-controllers and PCA for health screening. While this method improved cleanliness in public toilets, it faced issues related to medical test results.

Later Chandra et. al., developed a device auto calibration model using gas sensors and user feedback to create a hygiene monitoring system [4]. However, this system was expensive due to sensor calibration and manual inspection biases. [3] implemented an IoT-based public toilet management system along with occupancy monitoring using Arduino and ultrasonic sensors. Then, [23], also developed an IoT-based system for detecting gas and turbidity levels using Node-MCU micro-controller.

Anto et. al. and Parab et. al., implemented an automated cleaning system to improve public hygiene

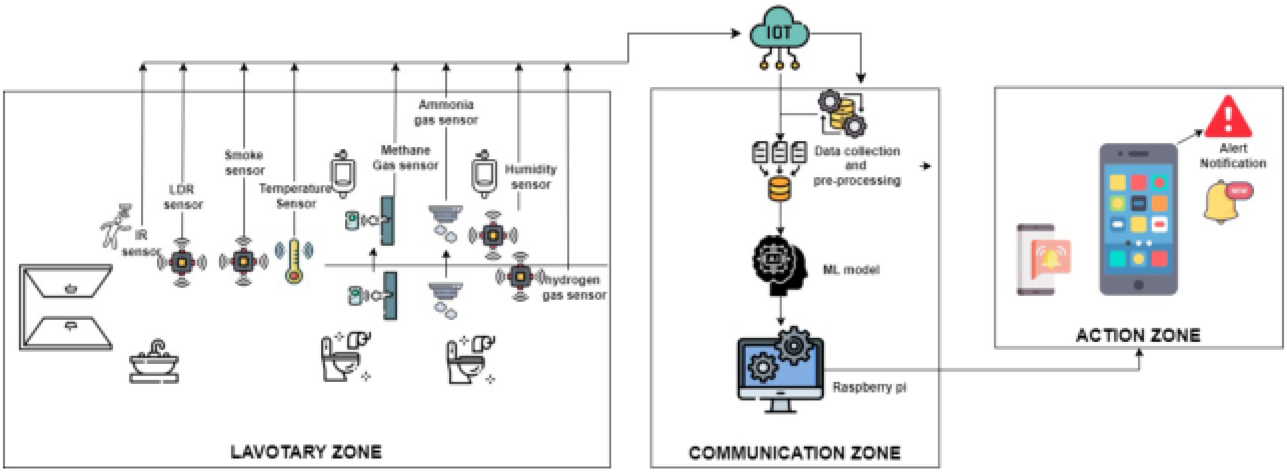


Fig. 3.1: Block Diagram of proposed system

but faced difficulties like resource unavailability, and complex screening mechanisms [1, 24]. Dhamale et. al. developed a system to monitor janitor activities and ensure real-time maintenance of public toilets but the high cost of a fully automated system is a major drawback [6].

Kadam et. al. and Mahalsekar et. al. have focused on providing contactless solutions and centralized monitoring systems to help prevent the spread of diseases caused by public toilets [10, 12]. According to [15], authors have proposed an automated monitoring and alert system for workers to clean the restrooms. Horadi et. al. have emphasized enhancing hygiene in public toilets using MQTT, HTTP, and predictive maintenance models [8].

The integration of IoT, ML, and DL in smart toilet systems has significantly improved public hygiene and resource efficiency. But, there are challenges such as sensor accuracy, costs, etc. For this, we implemented DL based approach(LSTM) for improving efficiency and hygiene. Model detailing is mentioned in a further section of the paper.

After reviewing the above different approaches we found that most of the papers were based on IoT and very few of them included the ML/DL approach. So in this paper, we have implemented the DL(LSTM) approach for improving the accuracy of the given system.

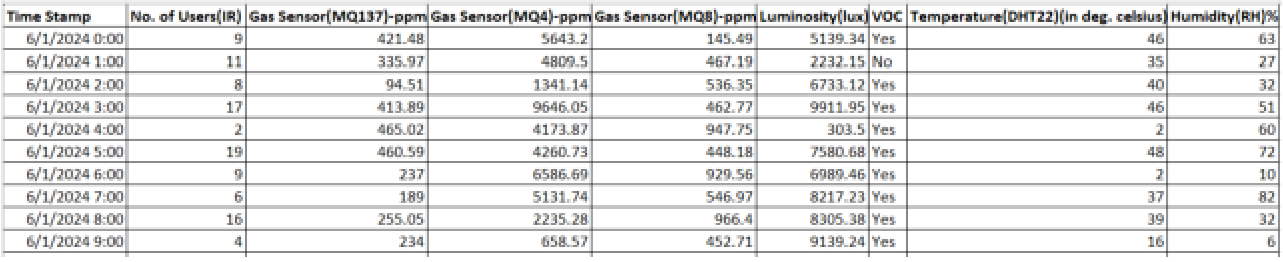
3. Proposed Model. A smart lavatory system is a modern approach to maintaining hygiene by collecting and analyzing sensor data, which is then processed by the DL model (LSTM) to predict whether the lavatory needs cleaning and thus send alerts to staff through the mobile application if required. As shown in Figure 3.1, The first zone is the lavatory zone, where sensors are placed, including a temperature sensor, ammonia gas sensor, VOC (smoke detection) sensor, LDR sensor, methane gas sensor, and IR sensor. Whenever a user enters the lavatory, each sensor provides different readings based on the user's activity. These readings are then combined and transmitted to the second zone, the communication zone, via the IoT device [13]. In the communication zone, data preprocessing is performed before it is fed into the DL model. The LSTM model is selected because the data is collected in real time, and LSTM is more suitable for continuous or sequential data, providing high accuracy. This prediction helps us determine whether cleanliness is required. To facilitate this process, we deployed the Raspberry Pi module. Once the decision is made, the control goes to the action zone where an alert/notification is sent to the staff for cleaning.

For our proposed approach we generate a dataset using various sensors. Figure 3.2 is a sample of the dataset used for the proposed approach. It consists first 10 records taken from sensor readings.

3.1. Real Time Dataset Description. We installed five sensors in each restroom, which contains two washbasins and six toilets. Each sensor was positioned between pairs of washbasins and toilets to monitor various environmental and usage conditions. During the six-month dataset collection period, hourly sensor

Table 2.1: Literature Survey

Ref.	**Objective**	**Methodology**	**Limitations**	**IOT**	**ML**	**DL**
Lokman et. al. 2017	To developing an IoT based smart toilet ensuring user privacy and implementing resource efficient scheduling algorithms	Improved efficiency, reliability and significantly decreased cleaner and energy costs by using genetic algorithms, ARIMA, kNN, and SVM models, as well as schedule optimization for predicting device duration.	• ARIMA model lacks RUL forecasting in specific settings. • Logarithmic transformation reduces time series data variation.	✓	✓	✓
Elavarasi et. al. 2018	Provide clean and hygienic toilet facilities to ensure public health and sanitation standards are met. and also contribute to the Vision of a 'Clean and Disease-Free India' through Innovative Technological Solutions	• Improved efficiency and cleanliness using microcontrollers for environment monitoring and PCA for health screening. • Smart toilets with sensors were implemented to monitor health and save water.	• There might be fault in the results of medical test	✓	✓	
Chandra et. al. 2018	Developing a device model auto-calibration by mapping user ratings to sensor readings and establishing a community benchmark hygiene rating system to enhance the user experience	• The Gas sensor and user ratings were combined to create a hygiene system for hygiene monitoring. • User can give feedback about odor in restroom using sensors installed in the area of restroom through bluetooth.	• Biases in manual inspection. • Sensor calibration for each installation is costly	✓	✓	
Cai et. al. 2019	Developing a system to enhance user experience in public toilets by monitoring toilet condition using IoT for cleanliness and other services like occupancy.	• Implemented Arduino with ultrasonic sensors for toilet monitoring and a Raspberry Pi for sanitation assessment. • IoT improved public toilet cleanliness, user experience and efficiency with a prototype system.	There is no limitations specified in this paper.	✓		
Sujeetha et. al. 2019	Implementing IoT-based sensors for effective toilet management system and spreading awareness about hygiene and maintenance of public toilet	• Sensors detect gas and turbidity levels in toilets for cleanliness. • Data is stored in firebase after processing by NodeMCU microcontroller.	There is no limitations specified in this paper.	✓		
Anto et. al. 2020	Enhancing user experience and janitorial tasks through automated the public hygiene.	• Water level indicator circuit • Automated flushing unit implementation with PIR sensor, servomotor and arduino UNO • Portable external unit for bowel cleaning without human effort.	• Manual cleaning by janitors is not reliable in public toilets. • Existing floor cleaning mechanisms are complex for washroom cubicles.	✓		
Parab et. al. 2020	Continuously monitoring the toilet conditions using sensors and implementing an automated toilet cleaning mechanisms for improving the hygiene.	• Ultrasonic sensor, gas sensor, Arduino controller, Microcontroller, LCD, buzzer, GSM, RFID reader for automatic monitoring system. • Engineering technologies for self-sustained E-toilet, sterilization, water-saving flushing mechanism.	• Unavailability of resources leads to unhygienic toilets. • Inadequate pay and safety equipment for maintenance staff.	✓	✓	
Dhamale et. al. 2020	Implementing a system to track the activities of janitor, thus ensuring real-time monitoring and maintenance of public toilets.	• Detecting unhygienic toilets using gas sensors like MQ-135. • Track worker activities using infrared sensors • Ultrasonic sensor at entrance to count number of toilet users.	• Fully automated systems are costly, and thus not feasible for all public toilets. • Monitoring in developing countries done manually, neglect in some regions.	✓		
Gong et. al. 2020	Research and analyze the development of smart toilets, focusing on their history, stages of development, and market penetration in different countries	• Smart toilet development divided into three stages: birth, growth, maturity. • Domestic smart toilet market growth from 2016 to 2020. • Quality of domestic smart toilet products improved from 2015 to 2019.	• Low penetration rate of domestic smart toilets, especially in lower-tier cities.			
Kadam et. al. 2021	Developing an IoT-based smart toilet and dustbin for hygiene and safety for implementing contactless solutions to prevent the spread of diseases.	Using efficient sensors for building E-toilets and E-dustbins using IoT technologies	• Accuracy of sensors decreases in sunlight, thus affecting their performance	✓		
Mahalsekar et. al. 2022	Developing a system to monitor public toilets for cleanliness and maintenance based on IoT	• Implementing an IoT based approach using five sensors namely RFID, MQ-135, IR, ultrasonic, and infrared for a centralized monitoring	• Lack of an organized approach to check the hygiene of public restrooms. • Insufficient maintenance and overpopulation lead to in insufficient public sanitation.	✓		
Chenchireddy et. al. 2022	Developing a system to monitor air quality by detecting harmful gases like CO2, smoke, and benzene and sound an alarm when they exceed a certain threshold . Also, the air quality will be displayed in PPM on LCD and website.	• IoT-based air pollution monitoring system with alarm for detecting harmful gases using sensors like MQ135, MQ6, and MQ2. • Serial UART, external interrupts, PWM, and SPI for data transmission	There is no limitations specified in this paper.	✓		
Patil et. al. 2023	Implementing an automated system that monitors and alerts janitors whenever the toilet conditions deteriorate past a certain threshold, thus revolutionizing restroom management practices	• IoT sensors will monitor ammonia, water levels, and motion in restrooms. • Data analytics will gove insights on water levels and occupancy. • A feedback monitoring system to enhance the user experience	• Lack of user interest in public sanitation affects cleanliness efforts. • Smart Toilet System focuses on air quality, odor control primarily.	✓	✓	
Horadi et. al. 2024	Enhancing public sanitation by improving cleanliness, hygiene and maintenance of public toilets through IoT-based smart toilet management system.	• System is designed with MQTT and HTTP for data transfer along with hardware implementation using Raspberry Pi and Arduino TTGO. • Implementation water-efficient automatic toilet flushing system using IoT technology by integrating sensors and predictive maintenance models for smart toilets. .	• Limited evaluation of the possibility and efficacy in the real world. • To confirm that the suggested approach is practical and efficient, further research is needed.	✓	✓	
Proposed approach 2024	• Improve Hygiene and resource efficiency. • Improve cleanliness and take care of human health.	• Estimate Clenliness LSTM model is used.	-	✓	✓	✓



Time Stamp	No. of Users(IR)	Gas Sensor(MQ137)-ppm	Gas Sensor(MQ4)-ppm	Gas Sensor(MQ8)-ppm	Luminosity(lux)	VOC	Temperature(DHT22)(in deg. celsius)	Humidity(RH)%
6/1/2024 0:00	9	421.48	5643.2	145.49	5139.34	Yes	46	63
6/1/2024 1:00	11	335.97	4809.5	467.19	2232.15	No	35	27
6/1/2024 2:00	8	94.51	1341.14	536.35	6733.12	Yes	40	32
6/1/2024 3:00	17	413.89	9646.05	462.77	9911.95	Yes	46	51
6/1/2024 4:00	2	465.02	4173.87	947.75	303.5	Yes	2	60
6/1/2024 5:00	19	460.59	4260.73	448.18	7580.68	Yes	48	72
6/1/2024 6:00	9	237	6586.69	929.56	6989.46	Yes	2	10
6/1/2024 7:00	6	189	5131.74	546.97	8217.23	Yes	37	82
6/1/2024 8:00	16	255.05	2235.28	966.4	8305.38	Yes	39	32
6/1/2024 9:00	4	234	658.57	452.71	9139.24	Yes	16	6

Fig. 3.2: Sample of the Dataset

data recordings were used to record important hygiene and usage pattern measurements. This model was implemented on a small scale using the proper sensors and setup equipment. By utilizing more advanced and scalable sensors, this system can be developed to monitor even more areas and cover larger areas. The description of each field and possible dataset values are given below.

1. Time Stamp: The objective of this feature is to track the timing of lavatory usage and environmental conditions, formatted as ”MM-DD-YYYY HH”. The time is updated on an hourly basis.
2. No. of Users (IR): This feature monitors lavatory occupancy to optimize cleaning schedules and improve overall hygiene management based on user density as detected by an infrared sensor. The values range from 1 to 20, reflecting fluctuations in lavatory usage.
3. Gas Sensor(MQ137)-ppm: The purpose of this feature is to detect ammonia levels helps in assessing air quality and control unpleasant odors, ensuring a healthier environment. The sensor detects values in the range of 5 ppm to 500 ppm.
4. Gas Sensor (MQ4) - ppm: This feature monitoring methane levels is critical for detecting leaks and ensuring safety in enclosed spaces, as well as maintaining proper air ventilation. The sensor detects values within the range of 100 ppm to 10,000 ppm.
5. Gas Sensor (MQ8) - ppm: The objective of this feature is to identify the presence of hydrogen, a potential safety hazard. The detection range spans from 100 ppm to 1,000 ppm.
6. Luminosity (lux): This feature captures the light intensity in the lavatory, measured in lux using an LDR sensor. The recorded values range from 5 lux to 10,000 lux, indicating varying lighting conditions. Maintaining optimal lighting ensures user comfort and energy efficiency, enabling smart lighting adjustments based on current conditions.
7. VOC (Volatile Organic Compounds): The purpose of this feature is to detect VOCs to indicate the presence of various volatile organic compounds, including cigarette smoke. It provides a binary value (yes/no). It helps in improving air quality and alerting users or staff for immediate action.
8. Temperature (DHT22) - °C: This feature records the temperature in the lavatory in degrees Celsius The detected values range from -40°C to 80°C. Monitoring temperature ensures comfort and can be used to regulate heating or cooling systems in the lavatory, contributing to energy efficiency.
9. Humidity (RH): This feature measures the relative humidity which is crucial for maintaining hygiene, preventing mold growth, and ensuring a pleasant user experience in enclosed spaces with values ranging from 0% to 100%.

3.2. Data analytics model. Long Short-Term Memory (LSTM) network is a specialized type of recurrent neural network (RNN), that is used for modeling long-term dependencies within sequential data. Due to its unique architecture, which can successfully capture temporal dynamics over long periods, it forms the basis in the fields of real-time data analysis and sequence prediction. It can update and maintain a memory cell, or internal state, during time steps. The structure of the architecture includes:

- Input Gate: It decides what new information should be added to the cell state. It has two parts: the input activation, for regulating the extent to which new information is stored, and the candidate cell state, for representing potential new information derived from the current input and previous hidden state.

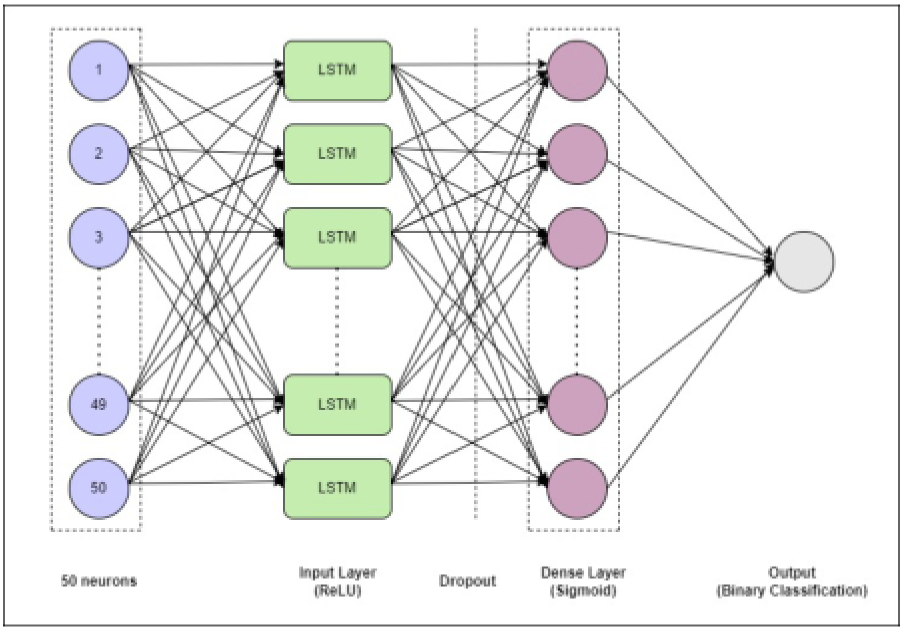


Fig. 3.3: LSTM Architecture

- Output Gate: This gate determines which parts of the cell state should be output as the hidden state for the current time step, contributing to the subsequent time step's computations.

We have implemented LSTM to predict the cleanliness of a lavatory based on data recorded using different sensors. The ability of LSTM to handle sequential dependencies was critical for capturing the dynamic environmental and usage patterns inherent in the data. The architecture included a single LSTM layer with 50 units and a relu activation function, followed by a dropout layer to prevent overfitting and a dense output layer with a sigmoid activation function to produce binary predictions, as shown in Figure 3.3.

3.3. Proposed model workflow. We developed a mobile application based on this proposed approach. In our initial scenario, a dataset is created by collecting readings from various sensors. After preprocessing, we generate the target value for "cleanliness," which determines whether the lavatory needs cleaning. We then apply the LSTM (Long Short-Term Memory) model to predict the target values and assess accuracy. After deploying the model to the cloud, it communicates with the mobile application, allowing workers to receive notifications that indicate when the lavatories need cleaning. This process relies on certain parameters, such as foul smell, temperature, and humidity. Workers are expected to clean the lavatories upon receiving these notifications.

According to this proposed method, a smart lavatory management system may be implemented, improving hygiene. Figure 3.4 illustrates how the model works.

4. Proposed Methodology. This section provides a detailed description of our proposed model in algorithmic form. Subsection 4.1 outlines the step-by-step procedure for generating the dataset using readings from various sensors. The subsequent section explains the data preprocessing process, including normalization and the prediction of the target class based on different recorded parameters. The final algorithm illustrates the training and evaluation process of the LSTM model. Subsection 4.4 details the different parameters such as the number of neurons, train-test split, and hyperparameters including the learning rate and activation functions.

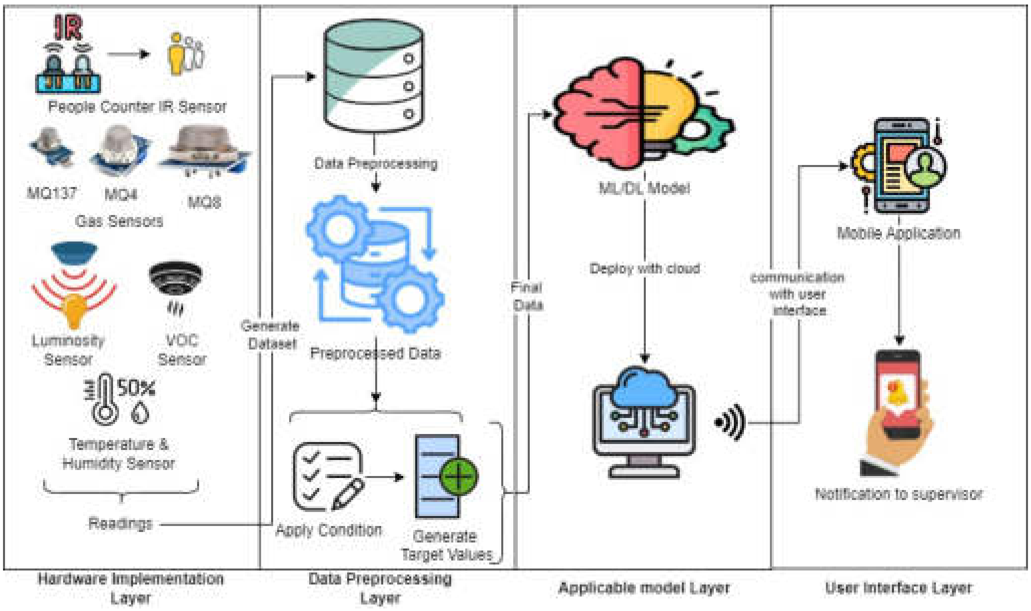


Fig. 3.4: System Architecture of proposed system

4.1. Dataset Generation. In this particular section, Algorithm 1 initializes various sensor readings and records data at regular intervals of one hour. There might be instances where the sensor might fail due to multiple reasons like inaccurate readings, which can result from various factors, including calibration problems, or external interference such as humidity, vibration, or extreme temperature. These issues can be resolved by regularly monitoring sensor behavior and using specialized sensors for extreme environments. The process is performed for a total of six months, appending each sensor reading to the dataset, ensuring an accurate and complete data set for analysis. As the algorithm works over the duration of the six months, it adapts to changes in conditions and sensor performance, which further improves the dataset's quality.

4.2. Preprocessing and Target class Prediction. After dataset preparation, z-score normalization is applied to the model as specified in Algorithm 2. The Z-score normalization process transforms the sensor values to a common scale to confirm that the data is standardized. Also, a new target class called ”cleanliness” is initialized, having a value of ’0’. Later, this target class changes according to the sensor parameters that were observed. The ’cleanliness’ target class is given a value of ’1’ specifically if any of the criteria exceed their set thresholds. The value ’0’ remains by the target class if all parameters remain within their threshold.

4.3. ML/DL Model. Algorithm 3 uses a Long-Short Term Memory (LSTM) model in the final step. The LSTM model is trained using the preprocessed sensor data to predict the target class. After training, the model's accuracy in predicting the ’cleanliness’ class is assessed, along with its overall accuracy. This evaluation helps identify when sensor data reveals cleanliness problems, enabling accurate and timely predictions.

4.4. Model's Parameters and Hyperparameters. The data is split with 20 percent reserved for testing and the remaining 80 percent for training. Each sample is treated as an individual time step with the number of timesteps set to 1. The sequential class is used to create a linear stack of layers to build the model layer by layer. Each layer contains 50 neurons, allowing it to detect dependencies over time. The activation function applied to the input of each LSTM unit is ReLU (Rectified Linear Unit). The dropout layer is a regularization technique used to prevent overfitting. It randomly sets 20 percent of the input units to 0 during each training update for better generalization. Only one output neuron is required, as it is a binary classification problem (clean or not clean) and using a sigmoid activation function to output a probability value 0 or 1. The Adam optimizer is suitable for large datasets and parameter sets, as it is an adaptive learning rate optimization

Algorithm 1 Generate Dataset

```
Input: All sensors, Time interval //(1 hour)
Output: Dataset with sensor readings
1. Initialize Sensors
  ammonia_value ← (input from MQ137 sensor) // Ammonia ($NH_3$)
  methane_value ← (input from MQ4 sensor) // Methane ($CH_4$) and Natural Gas
  hydrogen_value ← (input from MQ8 sensor) // Hydrogen ($H_2$)
  voc_value ← (input from VOC sensor) // Volatile Organic Compounds
  temp_value ← (input from temperature sensor) // Temperature
  humidity_value ← (input from humidity sensor) // Relative Humidity
2. For Each Time Interval, Collect Data
for t in range(0, total_duration, interval) do
    2.1 Get Current Timestamp
      timestamp ← get_current_time()
    2.2 Read Sensor Values
      occupancy ← read_IR()
      ammonia_value ← read_mq137()
      methane_value ← read_mq4()
      hydrogen_value ← read_mq8()
      light_intensity ← read_luminosity_sensor()
      voc_value ← read_voc_sensor()
      temp_value, humidity_value ← read_temp_humidity_sensor()
    2.3 Append Data to Dataset
      data_entry ← {
         'Timestamp': timestamp,
         'Occupancy': occupancy,
         'NH3 (ppm)': ammonia_value,
         'CH4 (ppm)': methane_value,
         'H2 (ppm)': hydrogen,
         'Light Intensity (lux)': light_intensity,
         'VOC (ppb)': voc_value,
         'Temp (°C)': temp_value,
         'Humidity (%)': humidity_value
      }
end for
```

technique that combines the best features of two existing stochastic gradient descent extensions: AdaGrad and RMSProp. Each time the model weights are updated, the learning rate determines the extent to which the model is adjusted in response to the estimated error. Adam typically employs a learning rate of 0.001 to balance the risk of overshooting the optimal response with the rate of convergence. For binary classification activities, binary Cross-Entropy Loss is used to measure the difference between true labels and expected probabilities.

5. Results and Discusions. This section briefly discusses the performance of the proposed approach using various evaluation measures. Features collected from different sensors were utilized to create the real-time dataset. Although the GRU (Gated Recurrent Unit) model was applied to the dataset, the results did not meet our expectations. The LSTM model performed better than the others on the available dataset. To train this model, we used ReLU and sigmoid activation functions, a learning rate of 0.001, and binary cross-entropy as the loss function (Equation 5.2). These parameters were chosen because our prediction model is designed for binary classification. The model is deployed on the Rasberry Pi module and communicates with a mobile device, sending alerts to the manager.

Different filtering techniques are applied to the model to determine its performance, with accuracy serving as the evaluation metric. In terms of binary classification, if y_i represents the true label for the i-th sample and

Algorithm 2 Predict Target class(Cleanliness)

```
Input: Dataset with sensor readings
Output: Preprocessed Dataset, Target class value (cleanliness)
1. Initialize Sensor Values
  ammonia_value ← (input from MQ137 sensor) // Ammonia ($NH_3$)
  methane_value ← (input from MQ4 sensor) // Methane ($CH_4$) and Natural Gas
  hydrogen_value ← (input from MQ8 sensor) // Hydrogen ($H_2$)
  voc_value ← (input from VOC sensor) // Volatile Organic Compounds
  temp_value ← (input from temperature sensor) // Temperature
  humidity_value ← (input from humidity sensor) // Relative Humidity
2. Initialize Threshold Values
  ammonia_threshold ← (define threshold for MQ137)
  methane_threshold ← (define threshold for MQ4)
  hydrogen_threshold ← (define threshold for MQ8)
  voc_threshold ← (define threshold for VOC)
  temp_threshold ← (define threshold for temperature)
  humidity_threshold ← (define threshold for humidity)
3. Initialize Cleanliness
  cleanliness ← 0
4. Preprocess Sensor Data
  4.1 Convert VOC Sensor Values
    (a) for all $i$ in range(len(df['VOC'])): do
          i. If df['VOC'][i] == 'Yes':
            A. df['VOC'][i] ← 1
          ii. Else:
            A. df['VOC'][i] ← 0
    (b) end for
  4.2 Standardize Sensor Values
    (a) sensor_columns ← ['NH3 (ppm)', 'CH4 (ppm)', 'H2 (ppm)', 'Light Intensity (lux)', 'Temp (°C)', 'Humidity (%)']
    (b) For all column in sensor_columns:
          i. $\mu_j \leftarrow \frac{1}{n}\sum_{i=1}^{n} X_{ij}$   // Mean of column $j$
          ii. $\sigma_j \leftarrow \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \mu_j)^2}$   // Standard deviation of column $j$
          iii. For all $i$ in range(len(df[column])):
            A. df[column][i] ← $\frac{\texttt{df[column][i]} - \mu_j}{\sigma_j}$
   5. Check Thresholds
   if ammonia_value > ammonia_threshold or methane_value > methane_threshold or hydrogen_value > hydrogen_threshold or voc_value > voc_threshold or temp_value > temp_threshold or humidity_value > humidity_threshold then
       cleanliness ← 1
   end if
   6. Output the Result
     OUTPUT cleanliness
```

$\hat{y}_i$ represents the predicted label, then:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \tag{5.1}$$

where N is the total number of samples.

The loss function typically used in binary classification tasks is binary cross-entropy. The formula for binary cross-entropy is:

Algorithm 3 Model Training

```
1: Input Preparation:
    1.1 Read normalized features
    1.2 Split dataset into train and test set
    1.3 Set hyperparameters (optimizer, learning rate, batch size)
2: Initialization:
    2.1 Set the number of epochs sufficiently large
3: LSTM Function:
4: function LSTM($x_t, e_{t-1}$)
    4.1 Local variables:
        (a) $i_t, o_t \in \mathbb{R}^N$
    4.2 Model weight matrices:
        (a) $W_i, W_o \in \mathbb{R}^{N\times M}$
    4.3 Model bias vector parameters:
        (a) $b_i, b_o \in \mathbb{R}^N$
    4.4 Compute gates:
        (a) $i_t = \text{relu}(W_i x_t + U_i e_{t-1} + b_i)$
        (b) $o_t = \text{sigmoid}(W_o x_t + U_o e_{t-1} + b_o)$
    4.5 return $e_t$
5: end function
6: Training Procedure:
    6.1 For each choice of neurons
        (a) For each range of number of replicates
              i. Train the model, monitor training loss
             ii. Repeat
                 A. Continue until validation loss at epoch $n \leq$ validation loss at epoch $n+1 <$ validation loss at epoch
                    $n+2$ (where n = Number of epochs)
                 B. or maximum epochs reached
            iii. Evaluate model on the test data
             iv. Calculate accuracy
        (b) Until validation loss criteria met
    6.2 End for
```

$$\text{Binary Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{5.2}$$

where y_i is the true label for the i-th sample (0 or 1), and $\hat{y}_i$ is the predicted probability for the i-th sample (output of the model's sigmoid function).

As shown in Table 2.1, the Adam optimizer trains the model using various epochs and activation functions. We analyzed the effectiveness of different combinations of activation functions and epoch counts in our experiments, aiming to determine the optimal configuration for enhancing the model's performance. The analysis revealed that training for 300 epochs with ReLU for the input layer and sigmoid for the output layer yielded the best performance compared to other configurations.

The relu activation function is a linear function that helps to avoid the vanishing gradient problem. This allows for faster and more efficient model training. It can also understand complex patterns. The sigmoid activation function works well for binary classification tasks, resulting in either 0 or 1.

The cross-validation technique can be applied to this dataset to assess how well the proposed model generalizes to an independent dataset. Using 20-fold cross-validation, 19 folds are used for training and 1 for testing. After training the model on all folds, we achieved a final avarage accuracy of 98.61%.

The stochastic gradient descent (SGD) optimizer is used to observe the trade-off in model performance (Table 5.2). This robust optimizer is effective in handling large data. However, its simplicity can sometimes be a disadvantage, as it may converge slowly.

Table 5.1: Accuracy with different activation functions and epochs for Adam optimizer

Test No.	Activation Function(i/p)	Activation Function(o/p)	Epochs	Optimizer	Accuracy
Initially	Relu	Sigmoid	100	Adam	-
1	Sigmoid	Relu	100	Adam	Decreased
2	Sigmoid	Sigmoid	100	Adam	Stable
3	Relu	Relu	100	Adam	Increased
Initially	Relu	Sigmoid	200	Adam	-
1	Sigmoid	Relu	200	Adam	Increased
2	Sigmoid	Sigmoid	200	Adam	Decreased
3	Relu	Relu	200	Adam	Increased
Initially	Relu	Sigmoid	300	Adam	-
1	Sigmoid	Relu	300	Adam	Stable
2	Sigmoid	Sigmoid	300	Adam	Stable
3	Relu	Relu	300	Adam	Stable

Table 5.2: Accuracy with different activation functions and epochs for SGD optimizer

Test No.	Activation Function(i/p)	Activation Function(o/p)	Epochs	Optimizer	Accuracy
Initially	Relu	Sigmoid	100	SGD	-
1	Sigmoid	Relu	100	SGD	Stable
2	Sigmoid	Sigmoid	100	SGD	Stable
3	Relu	Relu	100	SGD	Stable
Initially	Relu	Sigmoid	200	SGD	-
1	Sigmoid	Relu	200	SGD	Stable
2	Sigmoid	Sigmoid	200	SGD	Stable
3	Relu	Relu	200	SGD	Stable
Initially	Relu	Sigmoid	300	SGD	-
1	Sigmoid	Relu	300	SGD	Stable
2	Sigmoid	Sigmoid	300	SGD	Stable
3	Relu	Relu	300	SGD	Stable

Adam Optimizer has many advantages over SGD. One of the advantages is that it can maintain a dynamic learning rate. This results in faster convergence and improved performance. While SGD is a reliable optimizer, Adam's adaptive nature allows for better performance in complex models demonstrates that the choice of optimizer and the number of epochs significantly impact model performance. Thus, the optimal configuration for our proposed model is achieved using ReLU as the input layer activation function and combined with sigmoid for the output layer and the Adam optimizer. This setup resulted in a training accuracy of 99.83% and a test accuracy of 98.61%. The Table 5.3 presents a classification report, summarizing the performance metrics for the binary classification model.

The confusion matrix depicted in Figure 5.1 illustrates the performance of our binary classification model. The description of the figure is as follows. The model has correctly classified 6 instances as negative, misclassified 2 negative instances as positive and correctly classified 136 instances as positive and did not misclassify any positive instance as negative. As a result, the model achieved high accuracy by correctly predicting the class for the majority of instances. The precision for the positive class is 0.9855 while for the negative class is 1.0. The recall for the positive class is 1.0 and for the negative class is 0.75.

The matrix highlights the classifier's strength in accurately classifying positive instances, with only a few negative instances misclassified as positive. This type of visualization is crucial for evaluating the model's performance and identifying areas that require improvement.

Table 5.3: Classification Report

Class	Precision	Recall	F1-Score	Support
0	1.00	0.75	0.86	8
1	0.99	1.00	0.99	136
Accuracy	0.99			
Macro Avg	0.99	0.88	0.92	144
Weighted Avg	0.99	0.99	0.99	144

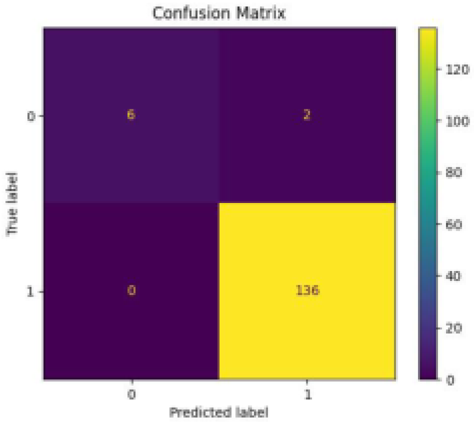


Fig. 5.1: Confusion Matrix of the Binary Classifier Demonstrating Model Performance

6. Conclusion and Future Directions. Smart Lavatory System combines IoT sensors and deep learning models to improve hygiene standards in public restrooms. The system effectively monitors lavatory hygiene conditions by continuously gathering data from environmental sensors. It guarantees timely cleaning and monitoring while offering extra features like occupancy sensors, air quality monitors, and automated dispensers to enhance the overall user experience. Deep Learning Algorithms such as LSTM and GRU, utilize real-time data that enable optimized allocation of cleaning staff and supplies. These automated system also help reduce the wastage of supplies such as hand wash, tissues. Future research will focus on scalability, advancing predictive capabilities, and developing user-friendly interfaces for real-time feedback. The authors aimed to improve the scalability of the existing framework by implementing data partitioning and sharing, distributing the data across multiple nodes and servers while accounting for all the washrooms on the university campus. This approach enhances query performance and simplifies the management of large datasets. The system's scalability is further demonstrated by its integration with a cloud infrastructure, offering elastic stability for handling high volumes of data. Additionally, the use of Deep Learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) further enhances its capability to handle extensive data processing efficiently.Data compression techniques can also be used to save storage costs and increase data transfer rates. Further, these datasets can additionally be processed effectively by implementing a scalable data processing framework like Apache Spark or MapReduce programming. Given that our current model has been designed for a limited workspace, we did not address data security or breach detection. In upcoming larger-scale implementations, we aim to implement RSA encryption to improve data security and defend against potential breaches. The Smart Lavatory Solution has the potential to keep evolving, delivering substantial advantages in public health, sanitation, and resource management, thereby raising hygiene standards.

REFERENCES

[1] I. M. Anto, B. Johnson, F. Wilson, A. Fitha, and A. Francis, *Arduino-based automated washroom sanitizing system*, in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 1196–1199.

[2] P. Banait, S. Kore, A. Shaikh, R. Marbate, D. Tayde, S. Katre, and A. Mahajan, *Automatic washroom cleaning system*, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, (2019), p. 37–39.

[3] W.-Z. Cai, N.-S. Chou, M.-F. Tsai, and Y.-C. Lin, *Intelligent toilet management system with internet of things technology*, in 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), 2019, pp. 1–2.

[4] S. Chandra, S. Srivastava, and A. Roy, *Public toilet hygiene monitoring and reporting system*, in 2018 IEEE SENSORS, IEEE, 2018, pp. 1–4.

[5] P. Deshmukh, A. Mohite, H. Bhoir, R. Patil, and A. Bhonde, *Intelligent public toilet monitoring system using iot*, in 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), 2020, pp. 1–6.

[6] V. Dhamale, S. Singh, S. Zadane, and M. Bhelande, *Smart toilet monitoring system using iot*, Department of Information Technology, Shah and Anchor Kutchhi Eng. College, Mumbai, India, (2020). Received Date: 13 February 2020, Revised Date: 25 March 2020, Accepted Date: 28 March 2020.

[7] M. K. Elavarasi, V. K. Suganthi, and J. Jayachitra, *Developing smart toilets using iot*, 2018.

[8] K. V. Horadi, M. S, and S. HL, *Smart public toilet management and monitoring system using iot*, International Journal of Advanced Research in Science, Communication and Technology, (2024), p. 344–353.

[9] hygiene-and health, *https://reports.essity.com/2018-19/hygiene-and-health-report/en/*.

[10] S. Kadam, B. Joshi, U. Gada, A. Chaugule, M. Bhelande, S. Rathod, and H. Motekar, *Iot based smart toilet and smart dustbin*, International Journal of Smart Home and Environment, 10 (2024), pp. 15–23.

[11] A. Lokman, R. K. Ramasamy, and C.-Y. Ting, *Scheduling and predictive maintenance for smart toilet*, IEEE Access, 11 (2023), pp. 17983–17999.

[12] N. P. Mahalsekar, S. Ganapaiah, R. R. Poojary, Sushmitha, and Sathisha, *Intelligent hygiene monitoring system for public toilets*, Electronics & Communication Engineering, (2024).

[13] M. Nayana, B. Chide, M. Nilesh, and P. Bobade, *Review: Iot based smart washroom*, International Research Journal of Engineering and Technology (IRJET), (2020), pp. 2090–2092.

[14] NIH, *https://www.nih.gov/*, Last Access- 25/06/2024.

[15] A. Patil, V. Poojary, S. Yadnik, U. Kokate, and R. Bhosale, *Towards the future of public restrooms: A smart toilet system for cleanliness and user satisfaction*, UG Student(B.Tech), (2024).

[16] T. R. Patil, B. N. Jayanthi, and A. C. Reddy, *Iot based energy efficacious smart hygiene system*, in 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021, pp. 272–277.

[17] Praneet Singavarapu and E. V. Murray, *Impact of inadequate sanitation on the marginalised*, (2013).

[18] restroom solution, *https://psiborg.in/smart-restroom-solution/*.

[19] D. Sasikala, V. V. Lakshmi, L. Jayanthi, S. Muthukumarasamy, D. Joy Winnie Wise, and P. Thirumaraiselvan, *Optimizing restroom resources: A smart toilet paper dispensing system using reinforcement learning algorithm*, in 2024 10th International Conference on Communication and Signal Processing (ICCSP), 2024, pp. 915–920.

[20] F. Shaikh, F. Shaikh, K. Sayed, N. Mittha, and N. Khan, *Smart toilet based on iot*, in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 248–250.

[21] smart-bathroom solutions, *https://www.faststreamtech.com/solutions/connected-smart-home-appliances/smart-bathroom-solutions/smart-restroom/*.

[22] smart-washroom-of-the future, *https://www.facilityapps.com/smart-washroom-of-the-future/*.

[23] R. Sujeetha, D. Abhinav, R. Rithik, and S. Abishek, *Toilet management system using iot*, in 2019 International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, 2019, pp. 783–787.

[24] T. A. P. V. T. B. R. R. G. Sushant A. Parab, Kashyap K. Meher, *E-swachh public toilet monitoring system*, International Research Journal of Engineering and Technology (IRJET).

[25] transparencymarketresearch, *https://www.transparencymarketresearch.com/smart-bathrooms-market.html*.

[26] uncleanbathroomdiseases, *https://enviro-master.com/commercial-cleaning-information/unclean-bathroom-diseases/*.

[27] S. Vardoulakis, D. A. Espinoza Oyarce, and E. Donner, *Transmission of covid-19 and other infectious diseases in public washrooms: A systematic review*, Science of The Total Environment, 803 (2022), p. 149932.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Jun 30, 2024
Accepted: Jan 7, 2025

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1071–1082, DOI 10.12694/scpe.v26i3.4134

INNOVATION-DRIVEN E-COMMERCE GROWTH, CONSTRAINTS, AND ADOPTION IN ORGANIZATIONAL PRACTICES IN THE 5G ERA

ABDULGHADER ABU REEMAH A ABDULLAH,[*] IBRAHIM MOHAMED [†], NURHIZAM SAFIE MOHD SATAR [‡], AND MOHAMMAD KAMRUL HASAN [§]

Abstract. Technology-driven e-commerce innovations have redefined how products and services are purchased and delivered online using the wireless 5G networks. With the transformative shift in e-commerce innovation, the commitment to use advanced 5G communication technologies to address e-commerce barriers has remained a determinant of organizational progress. To transform the landscape of e-commerce operations and improve the accessibility of products and services, this study explored the moderating effect of e-commerce barriers on organizational practices and e-commerce innovations based on the 5G network. To understand the impact of e-commerce barriers, 789 duly attested survey questionnaires were used to randomly collect samples from top management personnel from eighteen (18) actively functioning e-commerce companies across Libya. The statistical results show that e-commerce barriers influenced all the dimensions of organizational practices and the effort of e-commerce companies to adopt innovative practices. The coefficients of e-commerce barriers ($\beta = -0.223, t = -6.21, p < 0.05$) on various dimensions showed that e-commerce barriers are the main deterrents to digital transformation. Frequent updates and training on e-commerce innovations are recommended for organizations to align with the innovative trend in e-commerce developmental practices.

Key words: E-commerce big data, innovations and entrepreneurship practices, e-commerce barriers, e-commerce innovations in 5G networks.

1. Introduction.

Innovations in e-commerce (eCommerce) have been electronically driven to address constraints on sales of products and services in a virtual market. eCommerce has witnessed a remarkable shift, with an ever-increasing transformative innovation that enables customers to access various products online regardless of geographical location. Innovative changes in e-commerce have further enhanced customer transactions with features that support cutting-edge technologies that align organizational practices [41, 38]. For instance, the emergence of voice-driven commerce has created a retail market environment that enables customers to choose products that satisfy their needs [19, 46]. Innovations in e-commerce have evolved unprecedented marketing features that fundamentally support enterprises in shaping the retail market. The transformative innovations with 5G have increased service speeds and enhanced connectivity, which drive streamlined supply chains [47]. The improvement in e-commerce innovations is set to improve customer experiences.

Furthermore, the innovation of social commerce is rapidly transforming social media platforms into a potential virtual marketplace using blockchain to secure transactions of products and services over the Internet [51, 54, 53]. Social e-commerce enhances customer confidence in seamlessly initiating transactions. As e-commerce innovations change how products and services are delivered, different innovative features are frequently added to improve customer experience and increase the market size to accommodate more products and services [42, 52, 44, 35].

Organizational practices have evolved various strategies to improve a firm's operations and management [28, 3]. However, the workflow of practice and activities in organizations is driven by managerial philosophy and norms [34, 24]. Organizational practices represent the main structure that addresses eminent complex operations and long-term strategies [8]. [18] in a study noted that effective organizational methods supported and simplified management practices and ensured consistent development. Coherent organizational practices allow firms to

[*] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, UKM 43600, Malaysia (P99949@siswa.ukm.edu.my).

[†] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, UKM 43600, Malaysia

[‡] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, UKM 43600, Malaysia.

[§] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, UKM 43600, Malaysia.

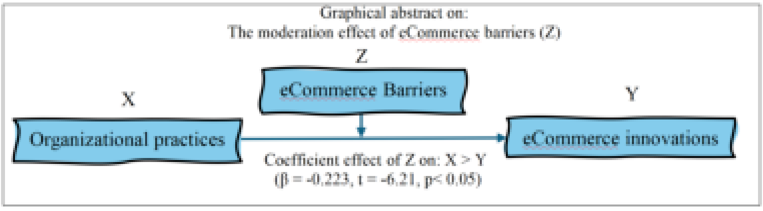

Graphical abstract on:
The moderation effect of eCommerce barriers (Z)
Z
eCommerce Barriers
X
Organizational practices
Y
eCommerce innovations
Coefficient effect of Z on: X > Y
(β = -0.223, t = -6.21, p< 0.05)

Fig. 1.1: The moderation effect of e-commerce barriers

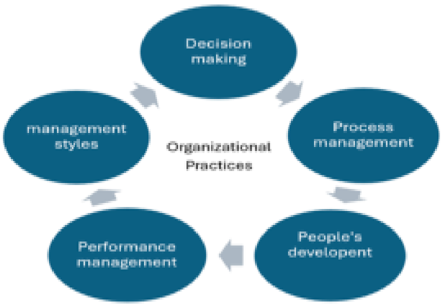


Fig. 2.1: Dimensions of organizational practices [7]

address barriers that could leverage opportunities to tap ever-transforming technology-driven business features to boost firm performance [5, 15].

Online e-commerce shops have opened new channels for firms and consumers to access various products and services. However, several challenges impede the smooth flow of business transactions and interaction with online customers [24, 49]. These challenges and obstacles are also referred to as barriers to e-commerce that prevent firms and consumers from benefiting. The range of barriers includes technological limits to regulatory complications and consumer trust concerns [33, 56, 43]. In developing countries, particularly Libya, the interaction between organizational practices, e-commerce developments, and e-commerce barriers plays a major role in the future competitiveness of commercial firms.

As e-commerce technology strives to align organizational practices with evolving customer demands, barriers associated with technology are inherent in online transactions. This study addresses the long-standing barriers to e-commerce innovation to establish a full digital economy. A clear understanding of the impact of e-commerce barriers on the complex interaction between organizational practices and e-commerce innovations could improve commercial activities in Libya and the economy.

2. Organizational Practices. Organizations systematically adopt a structured approach to manage daily activities and operations to achieve their goals effectively. The management processes that support an organization's overall functioning are called organizational practices [7]. Organizational practices improve over time and are based on collected experience, norms, and management philosophies contributing to achieving a specific goal. Organizational practices cover a range of activities guided by rules and regulations that shape the intended outcome [43, 22]. A study by [7] explained essential organizational practices in five dimensions: management style, decision-making, people's development, process management, and performance management, as shown in Fig. 2.1.

1. *Organizational practices vary across types and regions and are contextually defined by management priorities that align with the core values.*
2. Management practices are driven by organizational goals.
3. *Organizational practices were adopted by* [7] *and have been modified to explain eCommerce innovation based on the 5G network activities.*

4. *Conceptualized dimensions of organizational practices used in this research include management style, decision-making and people development, performance management, and process management.*

Management style refers to the specific approach managers or leaders use to execute their responsibilities and engage different teams to perform their duties within the organizational framework effectively. Decision-making involves picking a course of action or selecting options from numerous alternatives after carefully considering prospective outcomes. People's development is also referred to as employee development, which entails every deliberate action taken to improve the knowledge, skills, and potential of an organization's employees for a specific job. Process management systematically develops, regulates, and optimizes different activities and procedures to achieve organizational goals. Performance management is usually structured and planned and is achieved through careful monitoring, measurement, and improvement of the performance of individual workers or teams. Studies have shown that different activities constituting a firm's practices and operations contribute to human development and strengthen marketing strategies, customer service, and familiarity with technological features for online transactions [57, 16].

The organization's operations and activities are based on management decisions formulated and enforced to achieve specific goals of interest. Studies have shown that decision-making in e-commerce innovation is an important management role that fosters success and determines practices and resource allocation to improve innovation capability. [4] found that multi-criteria decision-making boosted customer investment in e-commerce across India. Decision-making based on an analytical fuzzy hierarchy process and fuzzy tops has been shown to play a key role in improving online shopping among Indians. The decision to invest in e-commerce innovation development has broadened the integration of e-commerce features into mobile and handheld devices with the help of the 5G network. The necessity to strengthen e-commerce innovation through decision-making forms the basis of the second research hypothesis, as follows:

Hypothesis 1 (H1): *Decision-making has a positive relationship with eCommerce innovations.*

The relevance of the e-commerce management style has focused on the activity processes that support online customers to initiate and complete a transaction [31, 48]. Studies have shown that e-commerce practices are strongly connected with emerging technology-driven innovations that widen the participation of different stakeholders and the size of products and services online.

Hypothesis 2 (H2): *Management style has a positive impact on eCommerce innovations.*

Innovation in e-commerce requires frequent updates, especially on new features that support online transactions. Users' knowledge and experience of new innovative features are important in promoting the sales of products and services. Studies have shown that the influences of people's development on e-commerce innovation affect its adoption, ease of use of innovative features, and the convenience of purchasing or selling products online [38, 21]. Knowledge of e-commerce business practices plays an important role in e-commerce innovation issues. The role of people's development has been demonstrated using a progressive model of cross-border e-commerce innovation and entrepreneurship. It was shown that e-commerce innovation was driven by people's digital development and required transformative reform to align with evolving technological innovations. The necessary knowledge and experience required for e-commerce form the basis of the third hypothesis of this study:

Hypothesis 3 (H3): *There is a positive relationship between people's development and eCommerce innovations.*

Process management supports the systematic execution and assessment of e-commerce innovation. This is because organizations' management processes follow specific procedures that support innovations. Studies have shown that the management of e-commerce business processes is required to make organizational practices more efficient and competitive. Another study [27] investigated the effectiveness and efficiency of management processes in e-commerce, considering quality-of-service delivery, employment, and customer retention using a model that focuses on previous literature. The findings show that management processes have a positive relationship with e-commerce innovation. Effective process management sustains innovation features and aligns with organizational practices' overarching strategic vision.

Hypothesis 4 (H4): *Process management has a positive impact on e-Commerce innovations.*

Performance management across employees, teams, and broader organizational entities directly impacts e-commerce innovation. Statistical results based on Data Envelope Analysis (DEA) and Stochastic Boundary

Analysis showed that performance management boosted e-commerce practices and promoted technological progress. The impact of performance facilitates easy access of innovation and organizational competitiveness. **Hypothesis 5 (H5):** *Performance management has a positive impact on e-commerce innovations.*

3. Ecommerce Innovations. Technological innovations have transformed the delivery of goods and services via online shops [57, 50]. For instance, integrating mobile devices such as smartphones and tablets into e-commerce platforms has increased the number of mobile shoppers (from 80 million in 2022 to 2.5 billion in 2023). It makes it convenient for potential online shoppers worldwide [37]. eCommerce innovations refer to novel and creative advancements, technologies, strategies, and practices that significantly improve and change how electronic business is conducted [45]. These innovations aim to enhance the online shopping experience, streamline business operations, and revolutionize the digital marketplace [55].

Innovative developments are constantly reshaping online commerce by opening new business opportunities and changing how consumers purchase and connect with brands not available in a local store. The emergence of social media platforms has made it easier to purchase products directly from online stores [53, 28]. In addition, augmented reality (AR) and virtual reality (VR) enable consumers to virtually visualize and interact with products before making a purchase; these features have added value to the e-commerce experience [16]. To further enlarge e-commerce investments, barriers to e-commerce innovations must be addressed to provide seamless online transactions that are more convenient at lower risk [20].

3.1. Innovations in Organizational Practices. The adoption of e-commerce has reduced the costs of products purchased online in Libya [40] and globally [30, 36, 39, 6]. eCommerce store presents a simplified product inventory of an organization in a format that can be accessed by online users. Organizations using eCommerce platforms do not need to purchase premises or physical space to store products and, by doing so, reduce initial costs. eCommerce has enlarged the global business platform for organizations in Libya, allowing them to provide products and services globally.

4. Ecommerce Barriers. eCommerce practices and innovations are influenced by organizational, technological, financial, and external constraints [21]. E-commerce barriers affect an organization's capacity to achieve a desired outcome. Technological barriers relate to outdated hardware or software that is compatible with innovation or difficult to integrate with new development or existing systems. Financial barriers refer to limitations arising from inadequate financial resources or that make it difficult for organizations to engage in e-commerce activities. External barriers arise from economic conditions, regulatory changes, market dynamics, technological advancements, cultural differences, and competitive pressure.

E-commerce barriers refer to factors that restrict the sales of products and services in organizations [7]. These barriers potentially influence organizational practices and the adoption of e-commerce innovations [14, 17, 2]. The success of e-commerce depends on the effort to address e-commerce barriers to unlock the possibilities of e-commerce benefits and to facilitate its widespread innovation [17]. Barriers to e-commerce can be observed at any stage in the organizational process, and its influence is more pronounced in online transactions, digital security, customer confidence, cross-border trade, and compliance with online regulatory requirements [12, 10]. **Hypothesis 6 (H6):** *eCommerce barriers harm the dimensions of organizational practices.*

4.1. Ecommerce Barriers in Organization. E-Commerce barriers constrain the advancement of commerce practices and reduce opportunities to widen their application. The success of organizational practices has been built around sales and productivity, which can only be effective and efficient in the digital age with e-commerce innovations. Barriers to e-commerce innovation affect organizational practices, development, and competitiveness.

E-commerce barriers negatively affect the seamless functionality and progress of e-commerce innovation [10]. A study on the effect of e-commerce barriers in Bangkok's small and medium-sized enterprises (SMEs) concluded that technology innovation prevented organizations from using e-commerce websites to sell products and services [9]. Other studies have shown that barriers to e-commerce progress significantly drop organizational development. The potential of management practices is drastically reduced by technological innovation, necessitating organizational planning to improve technological tools to enhance e-commerce adoption [13, 25, 58, 1].

The barrier to e-commerce requires urgent attention if enlarging the e-commerce market and enhancing

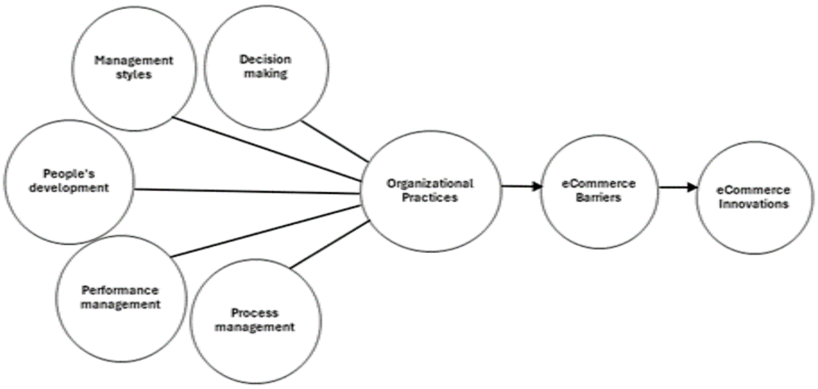


Fig. 4.1: The Research Framework.

organizational practices are interesting. Efforts to address e-commerce barriers can be translated into competitiveness, further improving income sources and organizational functionality.

Hypothesis 7 (H7): *E-Commerce barriers have a negative moderating impact on the relationship between organization practices and e-commerce innovations.*

The research framework of this study is shown in Fig. 4.1.

5. Research Method.

5.1. Partisipants and Procedures. The sample of this study comprised respondents from Libyan enterprises that were actively engaged in e-commerce practices. The selection process was on temporal engagement and included e-commerce employees with five to over 60 years of practical experience in e-commerce transactions in purchasing and selling products. These criteria for data collection were strategically implemented to secure reliable data and maintain the overall quality of the research findings. A total of 18 commercial companies that used e-commerce platforms across Libya were involved. The research target population focused on top employees with varying years of experience in playing different roles. Random sampling is appropriate for addressing biases from shared preferences and inclinations [45, 26]. valid research sample comprised 789 responses, which constituted 87.67 % of the total samples and was above 60% acceptable threshold for data analysis [11, 29].

5.2. Measures. E-Commerce barriers in the relationship between organizational practices and e-commerce innovations require substantial data to analyze the prevailing barriers quantitatively. Quantitative research based on a survey design was used to collect data from respondents. Participation in the survey was voluntary, and the research instrument was aligned to explore barriers to e-commerce innovation in Libyan organizations. Research ethics ensured confidentiality, response anonymity, and non-disclosure of information to third parties. The research instrument contained 54 items structured on a five-point 5.2.Likert scale ranging from strongly disagree to strongly agree.

6. Result and Discussion. The analysis presented in this section contained descriptive coefficients that summarized respondents' information relative to their knowledge of e-commerce practices and innovation across various organizations in Libya. Gender, age, tenure of service in e-commerce organizations, education, and management level were used to explain the respondents' views and perspectives relative to the study context. Statistical analysis shows that the dataset is characterized by experienced managers in various departments such as marketing, sales, services, products, regional managers, businesses, and research and development managers. The demographic distribution of presented in Table 6.1.

The means, standard deviations, and correlation statistics of the research variables are presented in Table 6.2. The central tendencies, denoted by the mean and standard deviation, explain the item variation. The

Table 6.1: Demographic profile of the respondents (N = 789)

Variables	**Category**	**Frequency (N)**	**Percentage (%)**
Gender	Male	531	67.3
	Female	258	32.7
Age structure	< 25	113	14.3
	25-35	270	34.2
	35-45	202	25.6
	45-55	162	20.6
	55-60	30	3.8
	> 60	12	1.5
Tenure of services	Less than < 1 year	99	12..5
	1-5 years	223	28.3
	5-10 years	249	31.5
	10-15 years	135	17.1
	15-20 years	58	7.4
	> 60	25	3.2
Level of education	Less than bachelor's degree	106	13.4
	Bachelor's degree	536	68.0
	Masters	130	16.5
	Doctorate	17	2.1
Management structure	Marketing manager	161	20.4
	Sales manager	168	21.3
	Service manager	128	16.2
	Product manager	110	13.9
	Regional manager	28	3.6
	Business manager	112	14.2
	Research and development manager.	82	10.4

interrelationships between variables were determined using correlation coefficients. A diagnostic measure was used to identify multicollinearity; the outcomes of the descriptive analyses are presented in Table 6.2.

Table 6.2 presents a detailed description of the variables using means and standard deviations. Analysis based on five Likert scale formats of the research items showed that the mean value for the dimensions of organizational practices was over 70 %.

Management style had a mean value of 3.60 (72 %) and was highly dispersed at SD = 0.89. Decision-making was 3.76 (75.2 %) and was highly dispersed from the mean value at SD = 0.73. Performance management had a mean value of 3.71 (74.2 %) and was well dispersed from the mean value at SD = 0.66. People's development had a mean value of 3.69 (73.8 %) and was dispersed from the mean value at SD = 0.67. Process management had a mean value of 3.64 (72.8 %) and was widely dispersed from the mean value of SD = 0.83. eCommerce barriers had a mean value of 3.62 (72.4 %) and were dispersed away from mean value at SD = 0.76. eCommerce innovations with a mean value of 3.41 (68.2 %) and were dispersed from the mean value at SD = 0.84. The results showed a moderate positive correlation between the e-commerce barrier, dimensions of organizational practices and e-commerce innovations ranging from r = 0.41 to r = 0.61.

6.1. Measurement Model. The moderated role of e-commerce barriers on the direct relationship between organizational practice dimensions and e-commerce innovations was evaluated. The statistical results for compound reliability and discriminant and convergent validity based on confirmatory factor analysis (CFA) are presented in Table 6.3. The accuracy of the findings was used to generalize innovative e-commerce practices across organizations.

The validity and reliability analyses provided insight into the overall quality of the measurement outcomes on factor loadings ranging from 0.682 to 0.772. Factor loadings exceeding 0.50 indicated a significant relationship with latent constructs [23, 32]. Average variance (AVE) was used to validate the convergent validity of each

Table 6.2: Mean, standard deviation, and correlation

Factors	Mean	SD	1	2	3	4	5	6	7
MS	3.60	0.89	1						
DM	3.76	0.73	0.41**	1					
PD	3.70	0.67	0.51**	0.48**	1				
PM	3.64	0.83	0.50**	0.55**	0.54**	1			
PfM	3.71	0.66	0.47*	0.52*	0.54**	0.52*	1		
eCB	3.62	0.76	-0.32	-0.41*	-0.42*	-0.40*	-0.26	1	
eCI	3.41	0.85	0.61*	0.60*	0.60**	0.51**	0.58**	0.41**	1

Note: * $P < 0.01, P < 0.05$,**

Table 6.3: Validity and reliability of the research instrument

Variables	Factor Loading	AVE	MSV	CR
DM	0.740	0.548	0.405	0.859
MS	0.742	0.553	0.427	0.861
PD	0.741	0.551	0.440	0.880
PM	0.772	0.596	0.450	0.856
PfM	0.682	0.531	0.448	0.850
eCB	0.730	0.535	0.382	0.902
eCI	0.754	0.570	0.473	0.889

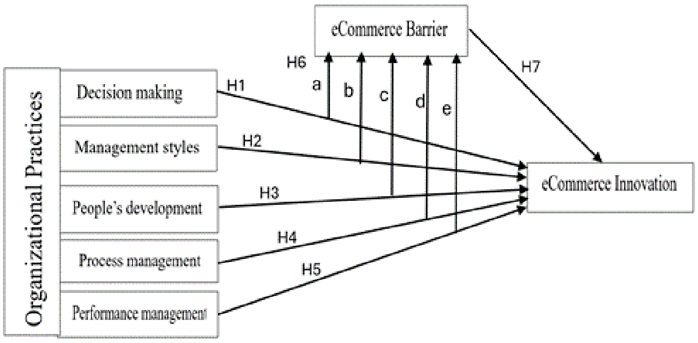


Fig. 6.1: Hypothetical path of the research variables

research factor. The two distinctive evaluation steps were compared between maximum shared variance (MSV) and AVE values and between correlation coefficients and the square root of AVE. The analysis showed that AVE values were superior to MSV. Composite reliability (CR) coefficients for the research variables ranged between 0.850 and 0.902, which exceeded the minimum threshold of 0.70 [32].

6.2. Hypothetical Path Analysis. The hypothetical path (H1 – H7) of the research hypothesis reported in this study is shown in Fig. 6.1. Different steps were adopted to provide an extensive evaluation of the relationships between the research variables. The relationship effect of the dimensions of organizational practices with e-commerce innovation (H1 – H5), as well as e-commerce barriers (H6) and the moderating effect of e-commerce barriers on e-commerce innovation (H7) were defined on the hypothetical path shown in Fig. 6.1.

Multilevel hierarchical regression (MHR) statistical analysis addressing the research hypotheses examined the effect of e-commerce barriers on the dimensions of organizational practices and e-commerce innovation. The influence of e-commerce barriers was classified into four steps representing a model of relationship across

Table 6.4: Validity and reliability of the research instrument

var	**Step 1**		**Step 2**		**Step 3**		**Step 4**	
	β	T	β	T	β	T	β	T
DM	0.49	12.07^{**}	0.51	13.27^{**}	0.52	14.11^{**}	-0.31	-763^{**}
MS	0.39	7.68^{*}	0.4	7.84^{*}	0.42	9.10^{**}	-0.15	-2.79
PD	0.36	6.39^{*}	0.38	6.73^{*}	0.38	8.04^{**}	-0.22	-4.71^{*}
PM	0.45	9.34^{**}	0.45	10.17^{***}	0.47	12.04^{***}	-0.19	-4.52^{*}
PfM	0.4	8.20^{*}	0.41	8.64^{**}	0.43	9.62^{**}	-0.14	-2.56
Moderation role of eCB								
eCB	-0.22	-5.92^{*}	-0.22	-6.21^{*}				
R^2	0.627							
Note: $^{*}P < 0.05,^{**} P < 0.01,^{***} P < 0.001$								

research factors, as shown in Table 6.4.

Table 6.4 is structured in four steps, with each measure representing a specific analysis that addresses the hypotheses of this study. The regression model results showed that the coefficient of decision-making varied ($\beta = 0.4880.522, t = 12.0714.11$ at $p < 0.001$), the coefficients of management style ranged ($\beta = 0.387 - 0.415, t = 7.68 - 9.10$ at $p < 0.01$), people development ($\beta = 0.362 - 0.380, t = 6.39 - 8.04$ at $p < 0.01$), process management varied ($\beta = 0.447 - 0.466, t = 9.34 - 12.04$ at $p < 0.001$) and performance management ($\beta = 0.399 - 0.425, t = 8.20 - 9.62$ at $p < 0.01$). The analysis showed varying degrees of impact of the dimensions of organizational practices, addressing hypotheses H1 − H5. The impact of e-commerce barriers on all dimensions of organizational practices (H6) was negative (presented in step 4). The model coefficient of e-commerce barriers on eCommerce innovation varied ($\beta = -0.219$ to $-0.223, t = -5.92$ to -6.21 at $p < 0.05$), which indicated a high adverse impact on eCommerce innovation which addressed the seventh hypothesis (H7). The statistical results show various negative effects of e-commerce barriers on the dimensions of organizational practices and e-commerce innovation. The adverse effects pose a significant limitation to e-commerce innovation based on the 5G network. Different models were classified under different steps to provide insight into varying influences of unresolved barriers. It is essential to acknowledge that incorporating the 5G network has significantly improved e-commerce services, considering high-speed communications and services as well as operational efficiency. However, organizations and businesses must consider the need for regular training employees, the high cost of 5G infrastructure, and regulatory challenges across locations legally supported by the services. This knowledge could enable organizations to plan and execute an effective and efficient business strategy properly. A summary of the findings on the benefits and constraints associated with e-commerce innovations based on the 5G network are presented in Table 6.5.

7. Conclusion and Recommendation. The present research successfully explored the impact of e-commerce barriers on organizational practices and e-commerce innovations in the 5G networks across firms in Libya. The performance of e-commerce organizations has demonstrated a commitment to adopting e-commerce innovation. Still, it has not earned a full dividend regarding e-commerce benefits because of the overriding e-commerce barriers. The 5G networks present significant opportunities for enhancing performance, customer experience, and operational efficiency of e-commerce. Therefore, prioritizing the innovativeness of e-commerce while addressing the barriers could open more opportunities to improve organizational competitive advantage. The data analysis reported in this study was well-structured to provide the depth of insight needed to clearly explain the differing barriers constraining e-commerce innovation at the organizational level. This finding can strengthen decisions on e-commerce innovations and practices to improve the customer experience. Preferences for a well-informed decision to improve employees' knowledge and skills on innovation and strategic practices to boost e-commerce practices are prerequisites for a successful e-commerce venture. Opportunities for continuous skill acquisition relative to e-commerce innovations will further empower employees to use diverse, innovative features to leverage competencies. Organizational leaders at the forefront of e-commerce innovation could be better positioned to handle e-commerce if they can assess the range of resources. Organizational practices must

Table 6.5: Research summary on 5G network innovations and e-commerce organization constraints

eCommerce Innovations using 5G Network	**Constraints using 5G Network**
High internet speed: The 5G network supports faster transactions and improves customer's shopping experience.	High cost of 5G infrastructures: Investments in 5G infrastructure require a huge amount of capital.
Rich Content: 5G seamless video streaming supports sales of products and services online and enhances advertisement of various contents.	Complexity with using the 5G network: Businesses require an upgrade to use innovative features of 5G such as augmented reality (AR) to enjoy faster services.
Improved customer experience: Online customers enjoy improved features and shopping experience of 5G based innovation.	Compatibility with older devices: Most widely used devices are not supported by 5G; limiting its adoption by most organizations.
Enhanced customer engagement: Fast communication and service using an advanced 5G network enable customers to engage in live chat shopping.	The risk with privacy and data security: Increasing the speed of data transmission tends to increase the vulnerabilities to cyber-attacks.
Real-time data processing: 5G enables e-commerce organizations to adopt real-time inventory tracking, and facilitate faster ordering of products and services.	Overloading of network: Congestion during peak times could slow down customer's access to service websites.
Efficient services: Faster network services with customers across different business operations support timely and efficient transactions.	Complexity with regulatory challenges: Incorporation of the 5G network strictly follows new regulations and environmental concerns that may complicate business compliance.
Improved decision and management services: Integration of 5G enables real-time tracking and optimization of management inventory and delivery speed of products and services.	Digital divide: The 5G network requires a certain level of sophisticated device that is not available to all consumers and as such, creates inequalities among users.
Enhanced growth of e-commerce innovations: E-commerce organizations can provide faster and more reliable services with a 5G network.	Constraints on service expectation: 5G network infrastructures are not available for customers in the less developed areas leaving service in some areas slower.

adopt a sound management style that supports the use of technology to foster efficient management operations and development endeavors. This decision could promote collaboration and support process management and employee performance. The 5G technology is set to play a significant role in maintaining efficient decision-making processes across various e-commerce organizations. Its capabilities in terms of service speed and access to data with connectivity to an automation system could contribute to making timely decisions and responding fast to online customer's needs. With the fast and flexible 5G network, management can quickly optimize various practices, adapt to situational changes to improve performance, and collaborate to equip online service teams better. To broaden the e-commerce experience, people's development will require urgent attention, especially in providing expatriates needed to improve the online sales of products and services.

The barriers relative to online shops affect potential customers' preference to purchase or order from e-commerce firms' services. A negative influence on organizational practices affects the advancement of e-commerce innovations. The impeding influence associated with organizational practices should be advocated to ease the integration of novel technologies and methodologies.

Valuable insights from the research findings could foster long-term e-commerce growth and support businesses in successfully adapting to dynamic digital markets by leveraging the sale of products and services across regional boundaries. In conclusion, e-commerce barriers have an unfavorable effect on organizational practices and adopting e-commerce innovation. The effectiveness of organizational practices drops with the emergence of barriers, as does their capability to fully harness the potential of e-commerce innovation.

Appropriate management practices should be formulated based on the findings of this study to address impeding barriers specifically. E-commerce strategies must be aligned with overarching organizational practices to address issues with technological innovations. To seamlessly utilize e-commerce opportunities, there is a need for collaboration and partnerships with technology providers and logistics companies and to improve payment methods. Training and skill development opportunities should be provided to improve and update e-commerce employees with emerging technological innovations and development trends. This will empower e-commerce workers to contribute immensely to e-commerce innovations and organizational progress. The findings of this study confirm that barriers to e-commerce affect different dimensions of organizational practices and reduce opportunities to sell products and offer services online. Top managers' knowledge of consumers requires close observation to understand how e-commerce barriers influence online business.

Management practices for the diversification of revenues are expected to combine corporate commitment and online selling approaches to ascribe the contribution of innovation to e-commerce progress. This study has successfully outlined the detrimental effect of e-commerce barriers on the development of online businesses. Therefore, the transformation of innovative e-commerce practices starts by addressing barriers that affect the use of innovative tools that support online business and, thus, e-commerce versatility. Managers can also use information and data from different sources to foster management processes and performance management. This will increase transparency and the efficient delivery of products and services based on the actual situation. The risk of merchandise running out of stock can be controlled by thoroughly evaluating e-commerce practices using information and data from different sources.

Acknowledgment. The authors acknowledge supporting funding from TAP & Faculty of Information Science and Technology Universiti Kebangsaan Malaysia. Dr Charles Ahamefula Ubani enriched and constructively added the paper to the presentation. The commitment of the editor-in-chief, associate editor, and reviewers' comments for improvement is among the efforts to put this paper in the best shape.

REFERENCES

[1] A. A. R. A. Abdullah, A. S. Madaki, I. Mohamed, N. S. Bin Mohd, and K. Ahmad, *The impact of it on knowledge sharing environment and management practice*, in 2022 International Conference on Cyber Resilience (ICCR), Oct. 2022, pp. 1–7.

[2] A. A. R. A. Abdullah, I. Mohamed, N. S. M. Satar, A. S. Madaki, and H. S. Hawedi, *Innovations in e-commerce development and the potential disruptive features*, in 2023 International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia, Oct. 2023, IEEE, pp. 1–6.

[3] I. Abualwafa, K. Ahmad, and U. Mokhtar, *A conceptual framework for knowledge management implementation in organizations*, Information Sciences Letters, 12 (2023), pp. 1547–1560.

[4] A. Agrawal, S. Bisht, M. Rathore, and A. Verma, *Application of multi criteria decision making in e-commerce sector*, in 2020 Research, Innovation, Knowledge Management and Technology Application for Business Sustainability (INBUSH), Feb. 2020, pp. 61–67.

[5] M. Akbari, M. E. Azbari, and M. H. Chaijani, *Performance of the firms in a free-trade zone: the role of institutional factors and resources*, European Management Review, 16 (2019), pp. 363–378.

[6] W. A. Ali, M. B. Mukhtar, and I. Mohamed, *The impact of technology, organizational, and trust factors on social commerce adoption*, Journal of theoretical and applied information technology, 97 (2019), pp. 2908–2921.

[7] Z. Ali, I. M. Zwetsloot, and N. Nada, *An empirical study to explore the interplay of Managerial and Operational capabilities to infuse organizational innovation in SMEs*, Procedia Computer Science, 158 (2019), pp. 260–269.

[8] A. Altrad, P. R. Pathmanathan, Y. Al Moaiad, Y. M. Endara, K. Aseh, Y. A. Baker El-Ebiary, M. Mohammed Farea, N. A. Abdul Latiff, and S. Iryani Ahmad Saany, *Amazon in business to customers and overcoming obstacles*, in 2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE), June 2021, pp. 175–179.

[9] Y. Amornkitvikai, S. Y. Tham, and J. Tangpoolcharoen, *Barriers and factors affecting e-commerce utilization of thai small and medium-sized enterprises in food and beverage and retail services*, Global Business Review, (2021), p. 097215092110362.

[10] K. Ariansyah, E. R. E. Sirait, B. A. Nugroho, and M. Suryanegara, *Drivers of and barriers to e-commerce adoption in Indonesia: Individuals' perspectives and the implications*, Telecommunications Policy, 45 (2021), p. 102219.

[11] T. Asparouhov, E. L. Hamaker, and B. Muthén, *Dynamic structural equation models*, Structural Equation Modeling: A Multidisciplinary Journal, 25 (2018), pp. 359–388.

[12] P. Aversa, S. Haefliger, F. Hueller, and D. G. Reza, *Customer complementarity in the digital space: Exploring Amazon's business model diversification*, Long Range Planning, 54 (2021), p. 101985.

[13] J. Ballerini, D. Herhausen, and A. Ferraris, *How commitment and platform adoption drive the e-commerce performance of SMEs: A mixed-method inquiry into e-commerce affordances*, International Journal of Information Management, 72 (2023), p. 102649.

[14] J. Ballerini, D. Yahiaoui, G. Giovando, and A. Ferraris, *E-commerce channel management on the manufacturers' side: ongoing debates and future research pathways*, Review of Managerial Science, (2023).

[15] L. Barros and I. Martínez-Zarzoso, *Systematic literature review on trade liberalization and sustainable development*, Sustainable Production and Consumption, 33 (2022), pp. 921–931.

[16] K. Baskaran and S. Rajavelu, *Digital innovation in industry 4. 0 era – rebooting uae's retail*, in 2020 International Conference on Communication and Signal Processing (ICCSP), July 2020, pp. 1614–1618.

[17] R. E. Bawack, E. Bonhoure, J.-R. K. Kamdjoug, and M. Giannakis, *How social media live streams affect online buyers: A uses and gratifications perspective*, International Journal of Information Management, 70 (2023), p. 102621.

[18] R. T. By and B. Burnes, eds., *Organizational change, leadership and ethics: leading organizations towards sustainability*, Routledge studies in organizational change & development, Routledge, Taylor & Francis Group, London ; New York, NY, 2nd edition ed., 2023.

[19] M. K. Chan and S. S. Kwok, *The pcdid approach: difference-in-differences when trends are potentially unparallel and stochastic*, Journal of Business & Economic Statistics, 40 (2022), pp. 1216–1233.

[20] K. Cheba, M. Kiba-Janiak, A. Baraniecka, and T. Kołakowski, *Impact of external factors on e-commerce market in cities and its implications on environment*, Sustainable Cities and Society, 72 (2021), p. 103032.

[21] M. Chen and R. Bashir, *Role of e-commerce and resource utilization for sustainable business development: goal of economic recovery after Covid-19*, Economic Change and Restructuring, 55 (2022), pp. 2663–2685.

[22] T. Christensen, P. Lægreid, and K. A. Røvik, *Organization theory and the public sector: instrument, culture and myth*, Routledge, Abingdon, Oxon ; New York, NY, second edition ed., 2020.

[23] J. M. Cortina, Z. Sheng, S. K. Keener, K. R. Keeler, L. K. Grubb, N. Schmitt, S. Tonidandel, K. M. Summerville, E. D. Heggestad, and G. C. Banks, *From alpha to omega and beyond! A look at the past, present, and (Possible) future of psychometric soundness in the Journal of Applied Psychology.*, Journal of Applied Psychology, 105 (2020), pp. 1351–1381.

[24] J. Eduardsen, S. Marinova, L. C. Leonidou, and P. Christodoulides, *Organizational influences and performance impact of cross-border e-commerce barriers: the moderating role of home country digital infrastructure and foreign market internet penetration*, Management International Review, 63 (2023), pp. 433–467.

[25] S. Elia, M. Giuffrida, M. M. Mariani, and S. Bresciani, *Resources and digital export: An RBV perspective on the role of digital technologies and capabilities in cross-border e-commerce*, Journal of Business Research, 132 (2021), pp. 158–169.

[26] M. Engert, J. Evers, A. Hein, and H. Krcmar, *The engagement of complementors and the role of platform boundary resources in e-commerce platform ecosystems*, Information Systems Frontiers, 24 (2022), pp. 2007–2025.

[27] Q. Farooq, P. Fu, Y. Hao, T. Jonathan, and Y. Zhang, *A review of management and importance of e-commerce implementation in service delivery of private express enterprises of china*, SAGE Open, 9 (2019), p. 215824401882419.

[28] L. L. Har, U. K. Rashid, L. T. Chuan, S. C. Sen, and L. Y. Xia, *Revolution of retail industry: from perspective of retail 1. 0 to 4. 0*, Procedia Computer Science, 200 (2022), pp. 1615–1625.

[29] M. Hecht and S. Zitzmann, *Sample size recommendations for continuous-time models: compensating shorter time series with larger numbers of persons and vice versa*, Structural Equation Modeling: A Multidisciplinary Journal, 28 (2021), pp. 229–236.

[30] Y. Inoue and M. Hashimoto, *Changes in consumer dynamics on general e-commerce platforms during the COVID-19 pandemic: An exploratory study of the Japanese market*, Heliyon, 8 (2022), p. e08867.

[31] M. Jara, D. Vyt, O. Mevel, T. Morvan, and N. Morvan, *Measuring customers benefits of click and collect*, Journal of Services Marketing, 32 (2018), pp. 430–442.

[32] O. P. John and V. Benet-Martínez, *Measurement: Reliability, construct validation, and scale construction*, in Handbook of research methods in social and personality psychology, 2nd ed, Cambridge University Press, New York, NY, US, 2014, pp. 473–503.

[33] S. U. Khan, A. Shah, and M. F. Rizwan, *Do financing constraints matter for technological and non-technological innovation? A (Re)examination of developing markets*, Emerging Markets Finance and Trade, 57 (2021), pp. 2739–2766.

[34] Y. Liu, C. Jiang, and H. Zhao, *Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media*, Decision Support Systems, 123 (2019), p. 113079.

[35] Y. Liu, M. Jin, S. Pan, C. Zhou, Y. Zheng, F. Xia, and P. S. Yu, *Graph self-supervised learning: a survey*, IEEE Transactions on Knowledge and Data Engineering, 35 (2023), pp. 5879–5900.

[36] Z. Liu and Z. Li, *A blockchain-based framework of cross-border e-commerce supply chain*, International Journal of Information Management, 52 (2020), p. 102059.

[37] A. Mosby, *99 online shopping statistics 2024 - total users & spending*, Feb. 2024.

[38] B. Niu, J. Dong, Z. Dai, and Y. Liu, *Sales data sharing to improve product development efficiency in cross-border e-commerce*, Electronic Commerce Research and Applications, 51 (2022), p. 101112.

[39] Oecd, *An oecd learning framework 2030*, in The Future of Education and Labor, G. Bast, E. G. Carayannis, and D. F. J. Campbell, eds., Springer International Publishing, Cham, 2019, pp. 23–35.

[40] H. F. H. Omar and M. M. Elmansori, *An empirical analysis investigating the adoption of e-commerce in Libyan small-medium enterprises*, International Journal of Business Information Systems, 37 (2021), p. 106.

[41] A. A. Peprah, C. Giachetti, M. M. Larsen, and T. S. Rajwani, *How business models evolve in weak institutional environments: the case of jumia, the amazon. Com of africa*, Organization Science, 33 (2022), pp. 431–463.

[42] ———, *How business models evolve in weak institutional environments: the case of jumia, the amazon. Com of africa*, Organization Science, 33 (2022), pp. 431–463.

[43] X. Qi, J. H. Chan, J. Hu, and Y. Li, *Motivations for selecting cross-border e-commerce as a foreign market entry mode*, Industrial Marketing Management, 89 (2020), pp. 50–60.

[44] D. Radcliffe, *Ecommerce in publishing: trends and strategies*, (2022).

[45] M. M. Rahman, M. I. Tabash, A. Salamzadeh, S. Abduli, and M. S. Rahaman, *Sampling techniques (Probability) for*

quantitative social science researchers: a conceptual guidelines with examples, SEEU Review, 17 (2022), pp. 42–51.

[46] P. Rao, N. S. Vihari, and S. S. Jabeen, *E-commerce and fashion retail industry: an empirical investigation on the online retail sector in the gulf cooperation council (Gcc) countries*, ICEB 2020 Proceedings (Hong Kong, SAR China), (2020).

[47] A. Rejeb and J. G. Keogh, *5G Networks in the Value Chain*, Wireless Personal Communications, 117 (2021), pp. 1577–1599.

[48] M. Saidani, H. Kim, N. Ayadhi, and B. Yannou, *Can online customer reviews help design more sustainable products? A preliminary study on amazon climate pledge friendly products*, in Volume 6: 33rd International Conference on Design Theory and Methodology (DTM), Virtual, Online, Aug. 2021, American Society of Mechanical Engineers, p. V006T06A002.

[49] X. Su, S. Wang, and F. Li, *The impact of digital transformation on esg performance based on the mediating effect of dynamic capabilities*, Sustainability, 15 (2023), p. 13506.

[50] Y. Tu and J. Z. Shangguan, *Cross-border e-commerce: a new driver of global trade*, in Emerging Issues in Global Marketing, J. Agarwal and T. Wu, eds., Springer International Publishing, Cham, 2018, pp. 93–117.

[51] I. Tzavlopoulos, K. Gotzamani, A. Andronikidis, and C. Vassiliadis, *Determining the impact of e-commerce quality on customers' perceived risk, satisfaction, value and loyalty*, International Journal of Quality and Service Sciences, 11 (2019), pp. 576–587.

[52] D. E. Ufua, O. J. Olujobi, and M. A. S. Al-Faryan, *Impact of the commercial law on e-commerce practices and business sustainability in nigeria*, International Journal of Management, Economics and Social Sciences, 12 (2023).

[53] D. E. Ufua, O. J. Olujobi, H. Tahir, V. Okafor, D. Imhonopi, and E. Osabuohien, *Social services provision and stakeholder engagement in the Nigerian informal sector: A systemic concept for transformation and business sustainability*, Business and Society Review, 127 (2022), pp. 403–421.

[54] S. Vakhariya, *A Study of Online Shopping Experience and Swaying Brand Preference Between Noon and Amazon in UAE*, South Asian Journal of Management, 27 (2020).

[55] Y. Vakulenko, P. Shams, D. Hellström, and K. Hjort, *Service innovation in e-commerce last mile delivery: Mapping the e-customer journey*, Journal of Business Research, 101 (2019), pp. 461–468.

[56] A. Valarezo, T. Perez-Amaral, T. Garin-Munoz, I. Herguera Garcia, and R. Lopez, *Drivers and barriers to cross-border e-commerce: Evidence from Spanish individual behavior*, Telecommunications Policy, 42 (2018), pp. 464–473.

[57] X. Zha, X. Zhang, Y. Liu, and B. Dan, *Bonded-warehouse or direct-mail? Logistics mode choice in a cross-border e-commerce supply chain with platform information sharing*, Electronic Commerce Research and Applications, 54 (2022), p. 101181.

[58] Y. Zhang, H. Long, L. Ma, S. Tu, Y. Li, and D. Ge, *Analysis of rural economic restructuring driven by e-commerce based on the space of flows: The case of Xiaying village in central China*, Journal of Rural Studies, 93 (2022), pp. 196–209.

Edited by: Ashish Bagwari
Special issue on: Adaptive AI-ML Technique for 6G/ Emerging Wireless Networks
Received: Jun 12, 2024
Accepted: Oct 3, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1083–1091, DOI 10.12694/scpe.v26i3.4375

COMPUTER NETWORK ATTACK DETECTION BASED ON JOINT CNN-LSTM MODEL WITH ATTENTION MECHANISM

MIAO JIANG*AND PEI LI†

Abstract. This paper addresses the problem that category imbalance in the traffic data set limits the detection performance of classification models for a few classes of attack traffic. The proposed method, which we call Jcla-detect, is based on a joint attention mechanism and a 1-D convolutional neural network (1DCNN)-Bi LSTM model. First, the Borderline SMOTE technique is used to pre-process the imbalanced training samples of traffic data during the data preparation step. This balances the various forms of traffic data and makes it possible for the subsequent model to correctly train the various types of data. After training a 1DCNN-BiLSTM model and a joint attention mechanism using the traffic data, the model extracts and classifies the local and long-range sequence characteristics. Then, by assigning a weight to the features that are helpful for categorization based on their significance, the attention mechanism raises the detection rate of the few assault types. The experimental results show that this method is effective in increasing the minority class attack traffic detection rate, as the method's detection accuracy can reach 93.17 for the URL dataset and it improves the detection rate of U2R attack traffic in the URL dataset by at least 13.70%.

Key words: Traffic anomaly detection; Category imbalance; CNN,Bi-LSTM; Attention mechanism

1. Introduction.

Web-based apps and services are becoming more and more important in people's lives as more and more network node devices are linked to the Internet. A significant number of network access devices—75 billion devices, to be exact—will be online by 2025, predicts Statista, a statistical research resource [1]. The flaws and vulnerabilities present in the protocols, operating systems, and application software that are employed in network attacks are constantly evolving and expanding in tandem with the substantial growth of the Internet. Traffic anomaly detection, an effective technique for network and information system security, is widely used to detect malicious behavior in network traffic [2].

Researchers have developed machine learning techniques to categorise and forecast massive amounts of traffic data for traffic anomaly detection as the amount of traffic data grows [3, 4]. A single classifier,it was discovered that typical machine learning techniques did not produce sufficient traffic anomaly detection results and that their detection performance was more feature-dependent. The majority of them prioritise feature engineering and feature selection, and thus frequently generate false alarms [5].

Numerous deep learning techniques have been used in recent years to study traffic anomaly detection by automatically extracting high-level features from the underlying traffic features through the neural network search space, with some promising research outcomes. To enhance the detection performance for NSL-KDD, [6] suggested a traffic anomaly detection approach combining stacked denoising self-encoder with soft max. [7] used densely linked CNN to detect traffic anomalies and increase detection precision using the KDDcup 99 dataset. On the CICIDS2017 dataset, [8] assessed the comparative detection performance of three neural networks, including LSTM, and discovered that Bi LSTM had the highest detection accuracy. In order to learn enough to improve detection outcomes, the majority of traffic anomaly detection approaches based on classical machine learning models and deep learning models need a lot of sample data. There is a notable class imbalance in the traffic statistics, with a considerable variation in the percentage of each attack class among the anomalous samples. Anomalies are typically greatly outnumbered by normal samples [9]. The majority class samples will outnumber the minority class samples when feeding this unbalanced traffic data training set straight into traditional classification models for learning and training, as is the case with less anomalous data and substantially unbalanced traffic data. Additionally, the few attacks with high threat levels may be

*Shangqiu Institute of Technology, School of Information and Electronic Engineering, HeNan Shangqiu 476000, China (jm20515@163.com).

†Shangqiu Institute of Technology, School of Information and Electronic Engineering, HeNan Shangqiu 476000, China.

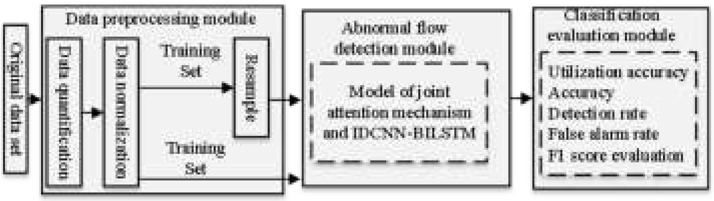


Fig. 2.1: Flow anomaly detection framework.

mistakenly identified as benign traffic or other attack classes, increasing risks to the network, devices, and users [10]. Therefore, the class imbalance issue in network traffic anomaly detection needs to be addressed in order to detect malicious activity in the network successfully.

Data and algorithms have been used by researchers to primarily address the category imbalance problem in traffic anomaly detection [11, 12]. On the data side, resampling techniques like synthetic minority class oversampling, adaptive synthetic sampling, and balanced resampling are mostly used to balance the different types of traffic data. On the algorithmic side, by enhancing algorithms or utilising integrated approaches, the detecting capability is increased. However, there is still a lot of space for improvement in the current research's detection rate of minority class attack traffic [13, 14]. This study suggests the Jcla-detect traffic anomaly detection method, which combines data improvement techniques with deep learning models to increase the detection rate of minority attack classes [15] to address the class imbalance problem in traffic anomaly detection.

The following are this paper's main contributions:

(1) In order to improve the detection performance of highly imbalanced traffic data in terms of both balanced data and improved models, this paper proposes a method for detecting traffic anomalies based on a joint attention mechanism and a 1DCNN-BiLSTM model.

(2) Using 1DCNN and Bi LSTM to extract local and long-range sequence features from network traffic data, respectively, we design a joint attention mechanism and a deep learning hybrid model of 1DCNN-BiLSTM for traffic anomaly detection in this paper. We also add an efficient attention mechanism to each block of 1DCNN and the end of Bi LSTM to focus on features that are crucial for classification and increase the detection rate of a few attack classes.

(3) This study conducts tests using the URL dataset and, using a number of evaluation criteria, compares the proposed method with several existing standard machine learning methods and methods that perform better on the problem of traffic data imbalance. The testing results demonstrate the method's superiority in the detection performance of imbalanced traffic data and its ability to greatly enhance the detection rate of a few classes of attack traffic.

2. The proposed model.

2.1. Traffic Anomaly Detection Framework. With a minimal number of harmful traffic samples, the traffic anomaly detection approach in this study aims to achieve excellent detection performance on traffic data. In this regard, the suggested data resampling methods and deep learning network models are combined in the traffic anomaly detection method. Figure 2.1 depicts the proposed method's overall detection structure, which is made up of three primary modules: data pre-processing, traffic anomaly detection, and classification and evaluation.

The data pre-processing module quantizes, normalizes, and resamples the original traffic feature data for training. Quantization and normalization allow the data to fit the deep learning model's input format specifications, while data resampling allows traffic data to be balanced and lessens the influence and bias of unbalanced initial data categories on the detection results.

In order to effectively detect a small amount of attack traffic with a high threat level, this paper develops a joint attention mechanism and a 1DCNN-BiLSTM model for deep traffic feature extraction and learning on the pre-processed traffic data.

The model's detection outcomes are reviewed and analysed in the classification evaluation module utilising a range of detection evaluation markers.

2.2. Data pre-processing.

(1) Measurement. Since the traffic feature data in this study includes non-numerical features that must be converted into numerical features, such as protocol type, service, and flag in the NSL-KDD dataset, the Label Encoder() function is used for label encoding. It is also necessary to convert the sample category labels into numerical form. For dichotomous classifications, the normal and assault labels are represented by 0 and1, respectively, and for multiple classifications, by a distinct heat for each sort of attack.

(2) Normalization. In order to reduce the size difference between the feature values in the flow data set, to avoid the impact of numerical magnitude differences and unit differences on the detection results, and to ensure that the detection results are valid, the Min-Max normalization method is used to map each feature data to the [0,1] interval, and its formula is shown in Eq2.1.

$$y_1 = \left\{ \; x_n = \frac{x - X_{\min}}{X_{\max} - X_{\min}} \right. \tag{2.1}$$

where x is each eigenvalue of the feature column X , $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the feature column X, respectively.

(3) Excessive sampling. When deep learning classifiers don't learn enough features during model training, it can make it harder for the model to identify specific kinds of assault samples. During the data preprocessing stage, it is required to oversample the minority class attack traffic so that the deep learning model may fully and efficiently understand the boundaries of each class in the traffic sample space. Although boundary samples are more important for generalization, misclassification is more likely to occur with them. Newly synthesized attack class samples need to be near the class boundaries in order to provide enough information for learning and detection. Using the borderline SMOTE approach, we oversample anomalous traffic samples in this investigation. We first locate the attack samples at the edges, then we regenerate the attack samples, and finally we add the freshly generated samples to the traffic data training set.

For each attack sample x in the training set, calculate its m nearest neighbors, if the number of normal samples in the nearest neighbors of x is more than the attack samples, then x as a boundary sample of the attack class is likely to be misclassified as a normal sample, and such boundary samples need to be oversampled. In the sampling process, the k nearest neighbor attack samples of the attack sample x are calculated, and $n(1 < n < k)$ attack samples are randomly selected from them. The formula for generating new samples of the attack class traffic is shown in equation (2):

$$y_1 = \left\{ \; T_n = T_i + rand(0,1) \times |T_j - T_i| \right. \tag{2.2}$$

where T_n is the newly generated sample, T_i is the boundary sample, T_j is the neighbor of T , and $rand(0,1)$ means generating a random number in the interval [0,1].

2.3. Traffic Detection Model. Since traffic data may be essentially thought of as sequence data with backward and forward correlation, traffic feature data, like the NSL-KDD dataset, exhibits significant correlation and backward and forward sequence dependency between separate features of the same sequence. The sequence learning model can be used to train the dataset in order to detect this type of assault by capturing the deeper features and correlation of the traffic data before and after the detection period. Over time, probe attacks could show up as a persistent change in traffic characteristics.

This study develops a traffic anomaly detection model with a joint attention mechanism and 1DCNN-BiLSTM to fully learn the traffic feature data and successfully extract its deep and complex features in order to increase the detection rate of minority attack traffic. 1DCNN is a good choice for sequence processing in this model since it can perform more non-linear transformation and give traffic sequence features a greater local feature learning power. Although 1DCNN has limited capability for long-distance learning models, the network traffic data has a time series structure and can classify recent long-distance connections based on earlier connections. Learning long distance sequence characteristics is the primary application of bi LSTM

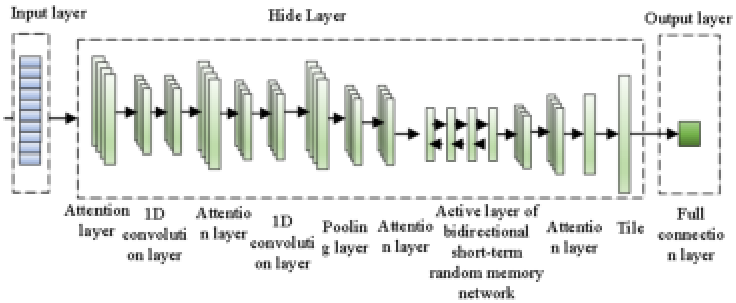


Fig. 2.2: Joint attention mechanism and 1DCNN-BiLSTM model structure.

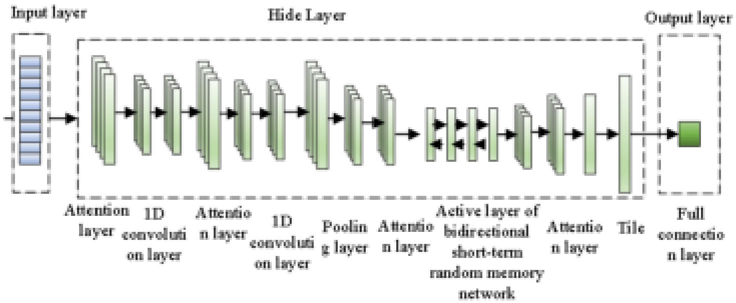


Fig. 2.3: 1 DCNN structure diagram.

networks. The collected deep traffic features from the 1DCNN are fed into the Bi LSTM to learn more sequential association patterns between deep traffic feature vectors across extended distances. In order to further enhance the model's performance in identifying unbalanced traffic data, this paper increases the weights of the traffic category-related features during the feature learning process. This causes the model to gravitate toward features that are more crucial for anomalous traffic detection.

To develop a multi-layer structure that includes a 1-dimensional convolutional layer, a pooling layer, an attention layer, a Bi-LSTM, a tiling layer, a fully connected layer, etc. to learn the correlation and local properties of normal and malicious traffic data sequences. Figure 2.2 depicts the structure of the model. Preprocessed traffic data is introduced into the model via the input layer, and the detection outcomes are then computed via the hidden layer and output via the output layer.

2.3.1. 1DCNN. For feature recognition, 1DCNN is a CNN that collects sequence data as a 1-dimensional grid. Despite having only one dimension, 1DCNN has the translation invariance of 2DCNN, which is advantageous for recognising features. Using this information as a foundation, this study generates the traffic feature data as sequence data with benign and malicious labels, and then applies 1DCNN to the traffic data to achieve local feature extraction. Figure 2.3 illustrates the structure of the 1DCNN model, which uses stacking of 1-dimensional convolutional and pooling layers to address the issue of local feature loss.

The first layer of convolution in one dimension is essential for feature extraction. By training the traffic data to produce an ideal set of convolution kernels with the least amount of loss, complicated traffic features can be automatically extracted using convolution kernels (filters). The i th sample of the flow data can be represented as an m dimensional feature vector $x_i \in R^m$, and multiple consecutive vectors $x_i, x_{i+1}, ..., x_j$ can be represented as $x_{i:j}$, 1-dimensional convolution is performed only in the vertical direction of the flow feature

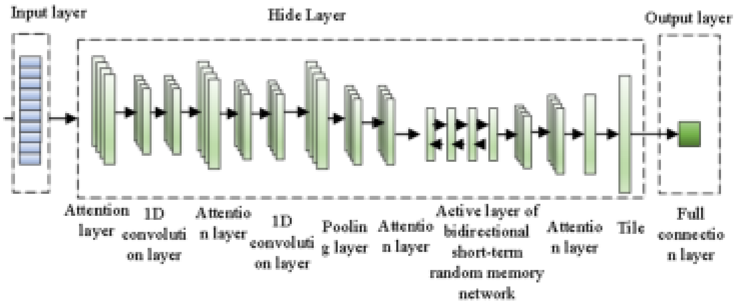


Fig. 2.4: Bi LSTM structure diagram.

data sequence, so the width of its convolution kernel is the dimension of the flow feature, and a feature mapping is constructed by applying a convolution operation to the input flow data using filter w to achieve local space feature extraction, which is calculated as Eq. (3) shows:

$$y_1 = \{ \ h_i = f(\omega \otimes x_{i:j} + b) \tag{2.3}$$

where b is the bias value and $f()$ denotes the nonlinear activation function linear rectification function for the convolution calculation. To retrieve the most crucial information, the pooling layer further aggregates and keeps the short-term features that the convolution layer extracted. The maximum pooling layer is employed in this study to merge the highest feature values from each convolutional layer's feature vectors. A 1 x n-dimensional data feature is created after the operation in the 1-dimensional convolutional and pooling layers, which effectively examines and preserves the local properties of the traffic data sequence.

2.3.2. Bi-LSTM. By learning the connection between the forward and reverse of sequence data, the Bi LSTM variant of the LSTM model improves it and gives it an advantage in classification tasks. The long-range sequence learning capabilities of the LSTM model are also present in Bi LSTM. The input traffic data is used in this study to train the forward and reverse LSTM of the Bi LSTM, whose structure is shown in Figure 2.4 and consists of an input layer, a forward hidden layer, a reverse hidden layer, and an output layer. On the other hand, the reverse LSTM retrieves the deep traffic feature sequence's reverse features from backward to forward. The input deep traffic feature sequence's forward properties are extracted by the forward LSTM. Both are combined in the output layer.

Bi LSTM effectively exploits the temporal features present in data before and after network traffic to improve model training, allowing the model to learn sequence features comprehensively.

2.3.3. The attention mechanism. The fundamental principle of the attention mechanism states that while irrelevant and useless information is ignored in favor of extracting features from more crucial and important information, limited attentional resources are allocated to a small number of crucial pieces of information that require special attention. In order to improve the detection rate of a small sample size of attack samples, it is more beneficial to implement an attention mechanism that assigns matching weights to different traffic features that are used to identify attacks when it comes to traffic anomaly detection. This paper introduces the bi LSTM network and the 1DCNN network, respectively, by means of the attention mechanism. The attention layer is added to the end of the convolutional block for the 1DCNN in order to solve the problem where the convolutional neural network only focuses on local characteristics and leads to erroneous learning of global information. The attention technique uses a weighted summing of its hidden layer vector output expressions to improve detection results with Bi LSTM. The attention mechanism uses probability to assign weights instead of the original random allocations.

3. Experimental results and analysis. On the URL dataset, Jcla-detect surpasses the machine learning techniques DT and LR. The results of Jcla-detect and the original CNN method are nearly identical. The

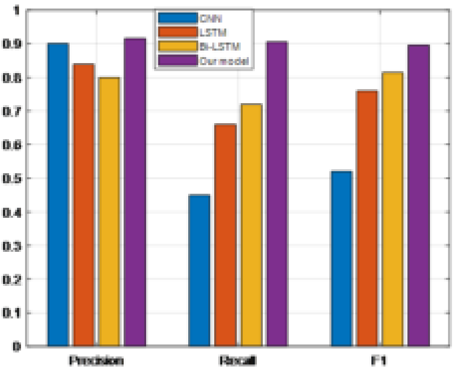


Fig. 3.1: F1 result of malicious URL data.

Table 3.1: Experimental results on TREC and 20NG.

Method	TREC	20NG	Average
SA-CNN	94.20	83.41	88.796
CNN	93.61	82.57	88.081
NB	90.41	82.78	86.586
KNN	86.53	75.29	80.901
LR	95.37	80.93	88.141
RF	90.73	73.45	82.081
DT	93.85	73.36	83.596
GBDT	93.89	67.58	80.726

proposed Jcla-detect is far more effective than the original CNN method in this regard because it can discover and visualise dangerous code sections.

3.1. Token segmentation and char segmentation. The experimental outcomes of the first CNN and Jcla-detect3 (LSTM) on the URL dataset are displayed in Figure 3.1. It would not be appropriate to utilise accuracy as an assessment metric at this time because malicious URLs typically make up a small portion of all URLs. Instead, we use F1 to assess the experimental findings for URLs. As can be seen, token segmentation produces better outcomes than char segmentation. This shows that the token segmentation method is successful in detecting URLs. Regarding the F1 assessment metric, Jcla-detect does not demonstrate a higher advantage over the original CNN approach, most likely because the performance of the original CNN itself is so high that it is very challenging to improve it.

The experimental results of Jcla-detect on two short text datasets are shown in Table 3.1, indicating that Jcla-detect outperforms the original CNN method.

We also use Bayesian, logistical regression, KNN, and GBDT methods to conduct experimental comparisons on the TREC and 20NG datasets. Jcla-detect performs better than these machine learning techniques, in our experience. This shows that the sequence attention technique works on both the short text dataset and the URL dataset.

3.2. LSTM model and Markov model. The Jcla-detect model itself was the focus of this paper's relevant research and experiments. The experimental Jcla-detect findings utilising LSTM and Markov language models on URL data sets in token partition mode are displayed in Figure 3.2. Compared to Jcla-detect-5 (LSTM) and Jcla-detect-3 (Mark-ov), Jcla-detect-3 (LSTM) is more suitable for anomaly detection utilising the token partition method, as shown in Figure 6. Jcla-detect-3 (LSTM) is suggested for URL detection tasks due to the URL length restriction and word correlation.

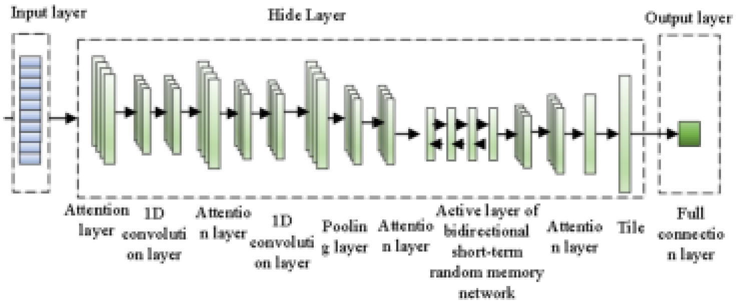


Fig. 3.2: Jcla-detect token partition results of three structures.

Table 3.2: Results of TREC and 20NG.

Methods	TREC	20NG
SA-CNN-3	94.20	83.409
SA-CNN-5	95.50	82.546
SA-CNN-M	93.42	80.457

3.3. Context length impactl. The experimental findings of Jcla-detect on two sets of brief text samples are presented in Table 3.2. TREC data set average length is 10, and 20NG average length is 5. Table Table 3.2 shows that Jcla-detect-5 (LSTM) outperforms competing approaches on TREC data, indicating that the average text length influences model choice.

The length of the URL is another potential factor that could influence the outcome of URL anomaly detection in addition to the token partition method previously mentioned. When the URL is brief, we might think about utilising Jcla-detect-3 (LSTM) to find anomalies. By examining Jcla-detect, we can observe that Jcla-detect-3 (LSTM) and Jcla-detect-5 (LSTM) are more suited for detecting URLs with longer lengths and that both models can produce more accurate results. Short URLs are unable to give additional information for anomaly identification, which further reduces the model's ability to detect anomalies. Therefore, we must take into account the length of the text while classifying texts or detecting anomalies, and then choose the right model to test.

3.4. Analysis of visualization results. For the sake of simplicity, this article provides some concrete examples to illustrate the results more intuitively. As shown in Figure 7, the darker the color is, the higher the value of attention is. By comparing the color depth, we can easily know which parts are malicious code areas. As shown in Figure 3.3, Local include file attacks attempt to access sensitive files on the server through the code "../". In article 3, char (106) is a statement to test whether the server can execute SQL functions. Because passwd, script, (;) and passwd are all offensive parts of malicious URLs, they all have high attention values. However, html and com11 have low attention values, because they are part of the normal code area. An interesting fact is that when the "." token is located in the normal URL, it has a low attention value. At this time, it is surrounded by oo4xccc and html, and the "." in the malicious code area has a high attention value. At this time, Its context is "/" and "./". This means that the context information generated by the external language model (LSTM/Markov) is valid and meaningful.

The overall effectiveness of several models on the CTU-13 data set is displayed in Table 3.3. The suggested model has better precision and recall rates than PCNN and HDM, and its training time is roughly a third of that of PCNN and HDCM. This is because the model suggested in this research can simultaneously learn tasks from three different domains, and training time is reduced through parameter sharing across domains.

The confidence interval shown in Table 3.4 was calculated using a paired sampling t test with a 90%

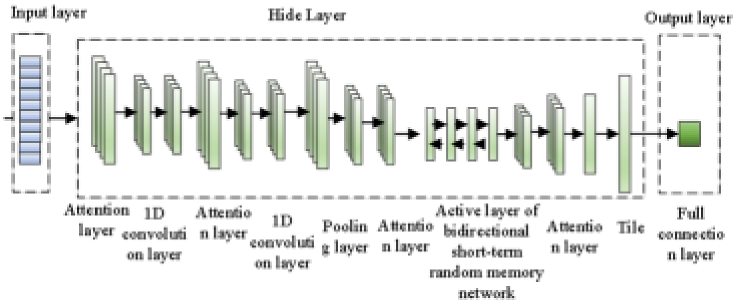


Fig. 3.3: Visualization result of malicious URL attention value of Chinese image title.

Table 3.3: Overall performance of models.

Measures	PCNN	HDCM	Proposed model
Precision (%)	97.5	94.8	99.4
Recall (%)	94.7	94.7	98.0
Sensitivity (%)	93.8	92.1	97.2
Accuracy (%)	95.6	93.4	99.1
Training time (s)	25.42	21.75	9.19

Table 3.4: Confidence interval.

-	$\mu_1 - \mu_3$	$\mu_2 - \mu_3$
Precision (%)	(3.561,1.316)	(6.711,0.985)
Recall (%)	(6.671,0.482)	(6.735,0.281)

confidence level. T stands for the average value of each measurement, and the subscripts 1, 2, and 3 denote the PCNN, HDM, and model that was proposed in this paper, respectively. There are 24 degrees of freedom (the abnormal category is 9 and there are 3 models), and the results were presented as a confidence interval. The findings demonstrate that the model put forward in this study has a higher recall rate than PCNN and HDM models and a higher precision than HDM models. The outcomes also demonstrate that the model put forward in this study is capable of handling three tasks from distinct domains simultaneously without degrading its performance in any one domain.

4. Conclusion. To enhance the detection rate of minority attack classes, we propose Jcla-detect, a novel method aimed at balancing the various types of traffic data. By addressing the inherent class imbalance, Jcla-detect ensures that the model is not biased towards majority traffic data, allowing it to learn more robust and accurate features from underrepresented attack traffic. The model undergoes comprehensive training across all traffic types, ensuring that both majority and minority classes are equally represented in the learning process.

One of the key innovations in our approach is the incorporation of a joint attention mechanism. This mechanism plays a critical role in refining the model's ability to focus on the most relevant features from different types of traffic data. By attending to both global and local patterns within the dataset, the joint attention mechanism enables the model to more effectively differentiate between benign and attack traffic, even in the presence of subtle variations. This enhanced feature extraction process is particularly beneficial for detecting minority attack traffic, which may exhibit less obvious characteristics compared to the majority class.

The experimental results demonstrate the significant impact of Jcla-detect in improving the detection rate of minority attack traffic. Specifically, our method was tested on the URL dataset, which contains both majority normal traffic and minority attack traffic. The findings indicate that Jcla-detect achieves a detection accuracy of 93.17%, outperforming traditional methods and highlighting the efficiency of our approach. This high accuracy is a direct result of the balanced training process and the joint attention mechanism, which allows the model to capture subtle features in attack traffic that may otherwise be overlooked.

Furthermore, Jcla-detect not only improves detection accuracy but also demonstrates greater robustness in dealing with complex and heterogeneous traffic data. The method effectively handles different traffic patterns, reducing the false positive rate and improving the overall reliability of the detection system. By balancing the data distribution and enhancing the feature learning process through attention mechanisms, Jcla-detect addresses key challenges in minority class detection, offering a promising solution for real-world applications where attack detection is critical.

In conclusion, the combination of data balancing and joint attention mechanisms in Jcla-detect significantly improves the detection of minority attack traffic, achieving high accuracy and robustness. These findings suggest that Jcla-detect can be a valuable tool in cybersecurity systems, particularly in environments with imbalanced traffic data. Future work could explore the adaptability of this approach to other datasets and attack types, further expanding its applicability and effectiveness.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

REFERENCES

[1] Hameed, K., Garg, S., Amin, M. B., Kang, B., Khan, A., *A context-aware information-based clone node attack detection scheme in internet of things.* Journal of network and computer applications (2022). 197.

[2] Weathersby, A., Washington, M., *Extracting network based attack narratives through use of the cyber kill chain: a replication study.* it - Information Technology, (2022). 64(1-2), 29-42.

[3] Tang, D., Wang, X., Yan, Y., Zhang, D., Zhao, H..*Adms: an online attack detection and mitigation system for ldos attacks via sdn.* Computer communications(Jan.), (2022).181.

[4] Ashiku, L., Dagli, C. , *Network intrusion detection system using deep learning.* Procedia Computer Science, (2021).185(1), 239-247.

[5] Venugopal, E., *A comparative analysis on hybrid svm for network intrusion detection system.* Turkish Journal of Computer and Mathematics Education (TURCOMAT), (2021). 12(2), 2674-2679.

[6] Winanta, A., Rochsantiningsih, D., Supriyadi, S. *Exploring efl classroom interaction: an analysis of teacher talk at senior high school level.* ELS Journal on Interdisciplinary Studies in Humanities, 3(3),(2020) 328-343.

[7] Sushma, E., *A review of the cluster based mobile adhoc network intrusion detection system.* Turkish Journal of Computer and Mathematics Education (TURCOMAT), (2021).12(2), 2070-2076.

[8] Han, S., Han, S., Liang, D., Liang, D., Haneda, M., Haneda, M. *A case study of two south korean middle school efl teachers' practices: instructional stances and use of classroom materials.* Classroom Discourse, 12(1-2),(2021) 56-74.

[9] Mao, B., Liu, J., Lai, Y., Sun, M. . *Mif: a multi-step attack scenario reconstruction and attack chains extraction method based on multi-information fusion.* Computer Networks(3), (2021). 108340.

[10] Lawal, M. A., Shaikh, R. A., Hassan, S. R. . *A ddos attack mitigation framework for iot networks using fog computing.* Procedia Computer Science, (2021).182(8), 13-20.

[11] Nandhakumar, E. . *A hybrid adaptive development algorithm and machine learning based method for intrusion detection and prevention system.* Turkish Journal of Computer and Mathematics Education (TURCOMAT),(2021). 12(5), 1226-1236.

[12] Nicolai, K. E.*A green gambit: the development of environmental foreign policy in morocco.* The Journal of North African Studies, 27(4), (2022) 714-740.

[13] Robinson, L., Brown, T., Gledhill, K., Isbel, S., Parsons, D., Etherington, J., et al.*'learning in and out of lockdown': a comparison of two groups of undergraduate occupational therapy students' engagement in online-only and blended education approaches during the covid-19 pandemic.* Australian Occupational Therapy Journal, 69(3), (2022) 301-315.

[14] Xiong, Y., Jin, M., Wang, J., & Wang, X. *Synergistic effect to enhance hydrogen generation of fe2o3/ce0.8sm0.1gd0.1o1.9 in water-gas shift with chemical looping.* International Journal of Energy Research, 46(7), (2022) 9733-9747.

[15] Hofmann, V., & C. M. Müller. *Challenging behaviour in students with intellectual disabilities: the role of individual and classmates' communication skills.* Journal of Intellectual Disability Research, 66(4), (2022) 353-367.

Edited by: Ashish Bagwari
Special issue on: Adaptive AI-ML Technique for 6G/ Emerging Wireless Networks
Received: Jul 25, 2024
Accepted: Oct 11, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1092–1103, DOI 10.12694/scpe.v26i3.4108

ENHANCING SECURITY OF CLOUD DATA USING CRYPTOGRAPHIC ALGORITHM BASED ON PFECCRS

AMRUTA GADAD*AND DEVI A†

Abstract. A web-based cloud computing application is basically used to save data with a view of accessing it from anywhere at any time. After analyzing the literature review, it is known that the work for cloud data security is either maintaining the security level or increasing the transmission speed of plain text of cloud, but failed to prove both security level as well as data transmission speed of cloud from one end to another end. Hence, to strengthen the data security of cloud and also to improve the data transmission speed, an integration of encoding, compression and cryptographic algorithms is important. An encoding technique of Prime Factorization (PF) for changing the original plain text into an intermediate plain text as encoded plain text followed by compression technique of Run Length Encoding (RLE) to reduce the file size so that the transmission speed of encoded message will be increased as well as the compression ratio will be higher and finally the Dynamic RSA algorithm is pertained to intensify the security by converting the compressed message into cipher text wherein Integrated Compressed Cryptosystem (ICC) and hence Prime Factorization Encoded Compressed Cryptosystems (PFECCRS) is proposed. The comparative analysis proved that the proposed methodology has increased the security level to 99.25%.

Key words: Prime Factorization, Encoding, Compression, Run Length Encoding, Encryption, Dynamic RSA.

1. Introduction. Cloud computing, a carriage of all computing assistance such as a carrier of software, servers majorly the databases, where each and every human try to save their data on this carriage. The security of this service carrier should be of prime concern to protect it from unauthorised users who may try to alter, destroy or misuse the data. The protection of all forms of cloud data can be done with the help of different concepts of cryptography, combination of compression and cryptography, or combining any mathematical encoding, compression and cryptography. The care must be taken that the data must be secured from several types of attacks such as phishing, replay attacks, cycle attack, fraudulent transactions, data stealing and many more [1]. A data is secured by converting the plain text into an unintelligible form of coded message and this process is called encryption. Transforming the cipher text back to original text is called decryption. The integrated process of encryption and decryption is known as cryptography. There are many cryptographic algorithms being used which are classified based on the type of key used. The usage of both public key and private key is known as asymmetric cryptography and only a single private key is said as symmetric cryptography. Symmetric-key encryption is the process where the plain text is converted into the non-readable text by using anyone of the symmetric-key encryption algorithm [2]. The converted non readable text is again decrypted back to the original plain text using the identical symmetric-key. Similarly asymmetric encryption is the process where the plain text is converted into the cipher text by using two separated keys basically known as public key and private key. The public key is used to convert the original plain text into the cipher text during the transfer from sender to receiver and private key is used during decryption i.e., converting the coded text message back to the original plain text [3].

Many different approaches are analysed to convert the plain text into encoded message using both public key and private key techniques, similarly there are different methodologies for compression, this compression helps in reducing the file size which furthers reduces the transmission speed and required storage space for file. There are mainly two approaches of compression lossless and lossy compression techniques. The lossless compression technique is best approached for text data and lossy works for image and other types of data. [4] Researchers have also showed how the different compression algorithms have also worked efficiently for cloud data[5].

*School of Computer Science and Applications, REVA University, Bangalore, India
†School of Computer Science and Applications, REVA University, Bangalore, India

Continuing further the document is structured into following sections. Background and related work are explained in section 2. Section 3 explains the relevant mathematical work used in this methodology. The proposed technique of Prime Factorization Encoded Compressed Cryptosystems (PFECCRS) is explained in section 4 and continued its illustration with an example in section 5. The experimental results of the same are discussed in section 6. Finally, section 7 ends up with conclusion.

2. Background and Related Work. The most widely used symmetric-key algorithms for data security are the stream cipher and block cipher algorithms. A stream cipher typically works on smaller units of plaintext, usually bits or bytes, whereas a block cipher symmetric-key algorithm converts a fixed length block of plaintext data into a block of ciphertext data of the same length The authors proposed the encoded compressed cryptosystems to improve the security level along with encoding using the Lucas and Fibonacci number systems and proved the security level to be 94.4% for 16MB files after Huffman compression and dynamic RSA [6]. The most commonly and strongly used symmetric algorithms are Advanced Encryption Standards (AES), Data Encryption Standards (DES) and many more [7,8]. Similarly, the asymmetric algorithms that are frequently used are Rivest-Shamir-Adleman (RSA), Elliptic Curve Cryptography (ECC) and some more [9]. The major functions used to analyse the strength of all the cryptographic algorithms are confidentiality, data integrity, security, authentication and non-repudiation.

As cryptography plays a vital role in data security similarly compression algorithms are also used to increase the transmission speed from one end to other end and then the encoding methods are used to convert the plain text into intermediate plaintext, to protect data from the hackers. The different compression algorithms that are widely used for data security are Huffman coding, RLE, Arithmetic encoding, Burrows wheeler transform (BWT) and many other algorithms resulting in good compression ratio by reducing the storage space and increasing the transmission speed [3]. These algorithms are classified as lossless and lossy data compression algorithms. The intermediate plain text can also be formed by different encoding algorithms such as Binary Number Systems, Fibonacci Series Lucas Series, two dimensional matrices and many others which helped to increase the security level of all the designed methodology [6,7,10].

Wid Akeel Awadh, Ali Salah Alasady, Mohammed S Hashim [13] proposed a multilayer data security model where the authors concentrated on merging the cryptographic and compression algorithm and further added a steganographic approach to enhance the security. AES-256 using RSA for encryption followed by Brotli compression and finally the LSB steganography technique ensured to achieve confidentiality, privacy, and integrity of the data. Sunday Adeola Ajagbe, Oluwashola David Adeniji, Adedayo Amos Olayiwola, Seun Femi Abiona [14] here the authors focused on AES based text encryption for NFC using Huffman Compression algorithm. AES was implemented in both ECB and CBC cipher-modes to compare performance, focusing on the time required for encryption. They mainly concentrated on implementing intrusion mitigation system to prevent interference in communication levels and integration with other security measures like multi-factor authentication for fortification. Shiladitya Bhattacharjee, Himanshi Sharma, Tanupriya Choudhury, Ahmed M. Abdelmoniem [15] the authors proposed a combined approach to enhancing encryption and compression algorithms for large data transfer. The chaotic S box encryption and adaptive Huffman compression algorithm proved to achieve superior time and space efficiency with enhanced privacy and integrity for any generic data in terms of entropy, bits per code, information loss percentage, and throughput.

N. Sugirtham, R. Sherine Jenny, B. Thiyaneswaran, S. Kumarganesh, C. Venkatesan, K. Martin Sagayam, Lam Dang, Linh Dinh, Hien Dang, [16] explained using a modified Playfair algorithm, partitioning the plaintext, adding filler characters, inserting filler information, compressing using LZMA, and utilizing a variety of encoding schemes are all part of the methodology. The suggested approach removes fillers for authentic retrieval and fortifies the Playfair cipher. With only minor key changes, the avalanche effect ranges from 65% to 93.7%. For compressed, secure text, the encrypted document is further encrypted using LZMA. The complete study of all such different cryptographic and compression algorithms used for data security are as explained, which says how each cryptographic algorithm merged or unmerged with compression algorithms are how efficient in satisfying any of the parameters like efficiency, security level, integrity and many more [9].

3. Mathematical Background.

Prime Factorization. Let $a_m(n)$ be the sum of the m^{th} powers of the primes in the prime factorization of n. For example, $a1(2^3.5.11^3)$ = 2+2+2+5+11+11+11, $a2(2^3.5.11^3)$ =22+22+22+52+112+112+112, $a5(2^3.5.11^3)$

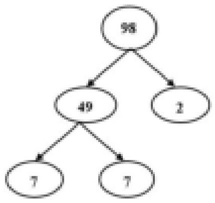


Fig. 3.1: Example for the process of prime factorization

= 25+25+25+55+115+115+115. Let $b_m(n)$ be the mth power of the maximum prime factor in the prime factorization of n. For example, $b1(2^3.5.11^3)$ = 11, $b2(2^3.5.11^3)$ = 112, $b5(2^3.5.11^3)$ = 11^5 [9].
The prime factors of any non-prime integer n can be found among a set.

$$(P1, P2, ..Pk)\ where\ \ Pi \leq \sqrt{n}, 1 \leq i \leq n,$$

where P1, P2,..., Pk are the prime factors that the trivial division method finds for the given number n. This trial division method, divides n by smaller prime numbers (beginning with 2, 3, 5, 7 and so on) in a blind manner and is the simplest way to factor n. If the remainder of division is zero, a prime number is chosen as a factor. This process is continued until all prime numbers that are less than or equal to n are identified and hence is used to factor small integers formed by some digits, but it is not suitable for large numbers due to its enormous time complexity [12].

For example, calculating the prime factors for the ASCII value of letter A which is 98 and hence the value of n = 98. i.e., 98 = 49 x 2, as shown in the Fig. 3.1. The factors found for the number 98 by using trivial division method where the number 98 is divided by the smallest possible prime number 2, in the second step the value 49 is processed through trivial division and hence dividing it by 7 times resulting as 7 x 7. The final obtained prime factors for the number 98 is $2(7^2)$.

4. Proposed Methodology. The proposed methodology is illustrated in Fig. 4.1 which is basically designed to strengthen the data security and increase the transmission speed of plain text during the transfer of data from one end to another end. The plain text PT of cloud server is initially converted into intermediate Encoded Plain Text (EPT) before it is encrypted. The EPT is generated by applying prime factorization for all ASCII values of the plain text, these prime factors are converted into binary digits which intern forms the first level of data security. The EPT is not encrypted directly instead it is processed through the RLE Compression algorithm first forming the next intermediate message known as Intermediate Compressed Message (ICM), this ICM helps in strengthening the rate of data transfer from the user to cloud server. The ICM is finally used in forming the cipher text applying Dynamic RSA where the intermediate message ICM is given as input message for RSA algorithm whose block size is less than n (formed from two distinct large prime numbers). This cipher text on the sender side is converted back to the plain text by reversing the process i.e., the encrypted cipher text is subjected to decryption of Dynamic RSA were finding back the message ICM which is again processed through decompression of RLE obtaining back the Intermediate Decompressed Message IDCM. The IDCM is decoded back to the original PT by using reverse process of prime factorization.

4.1. Prime Factorization. To encode the PT of cloud into intermediate EPT, the prime factors are found for the ASCII values of the plain text. The resultant prime factors are further substituted into equivalent binary numbers. The binary values formed for each ASCII value is considered and the process is repeated for all the letters of the given PT. After conversion of all ASCII values into prime factors followed by substitution into binary numbers, all these binary numbers are merged and found the intermediate EPT. The PT processing through multiple steps to find the intermediate EPT is as shown below in Algorithm 5.

4.2. Compression with Run Length Encoding. Compression techniques are basically classified as Lossy and Lossless compression methods. RLE is also one of the Lossless data compression techniques which is applied when data is the sequence of characters in which some particular characters are repeated consecutively

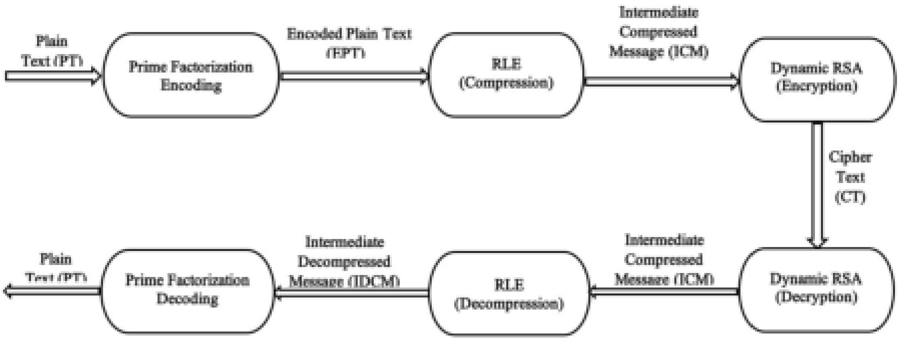
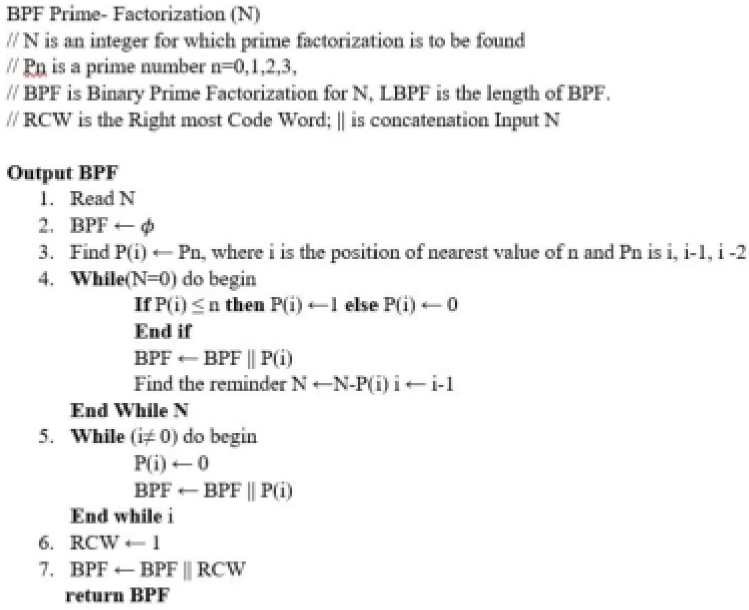

Plain Text (PT)
Prime Factorization Encoding
Encoded Plain Text (EPT)
RLE (Compression)
Intermediate Compressed Message (ICM)
Dynamic RSA (Encryption)
Cipher Text (CT)
Dynamic RSA (Decryption)
Intermediate Compressed Message (ICM)
RLE (Decompression)
Intermediate Decompressed Message (IDCM)
Prime Factorization Decoding
Plain Text (PT)

Fig. 4.1: Proposed PFECCRS scheme

Algorithm 5 Prime factorization



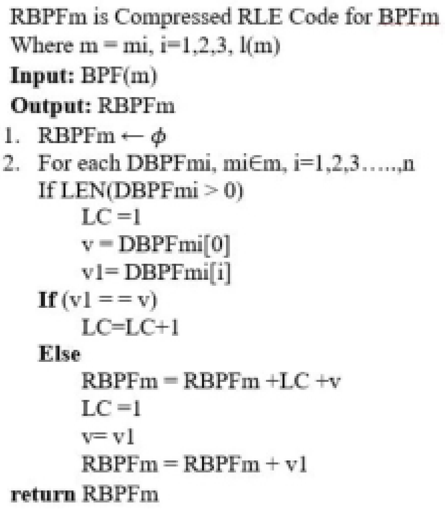
```
BPF Prime- Factorization (N)
// N is an integer for which prime factorization is to be found
// Pn is a prime number n=0,1,2,3,
// BPF is Binary Prime Factorization for N, LBPF is the length of BPF.
// RCW is the Right most Code Word; || is concatenation Input N

Output BPF
  1. Read N
  2. BPF ← ϕ
  3. Find P(i) ← Pn, where i is the position of nearest value of n and Pn is i, i-1, i -2
  4. While(N=0) do begin
         If P(i) ≤ n then P(i) ←1 else P(i) ← 0
         End if
         BPF ← BPF || P(i)
         Find the reminder N ←N-P(i) i ← i-1
     End While N
  5. While (i≠ 0) do begin
         P(i) ← 0
         BPF ← BPF || P(i)
     End while i
  6. RCW ← 1
  7. BPF ← BPF || RCW
     return BPF
```

many times. The consecutive repeated characters are compressed by representing it with a number which tells how many times the character has been repeated consecutively.

If the sequence is of the form WWWWWWWEEEETTTTTTTTTTTDDDDDSSSSSSS then this sequence can be compressed using RLE as 7W4E11T5D7S so that instead of occupying 34 bytes of memory it can be reduced to only 11 bytes of memory. This process of RLE is majorly used in image compression and binary sequence compression. As the EPT obtained after PF is a binary sequence of characters, the use of RLE could be justified and the same is elaborated in the algorithm 6 for the BPF to obtain the ICM.

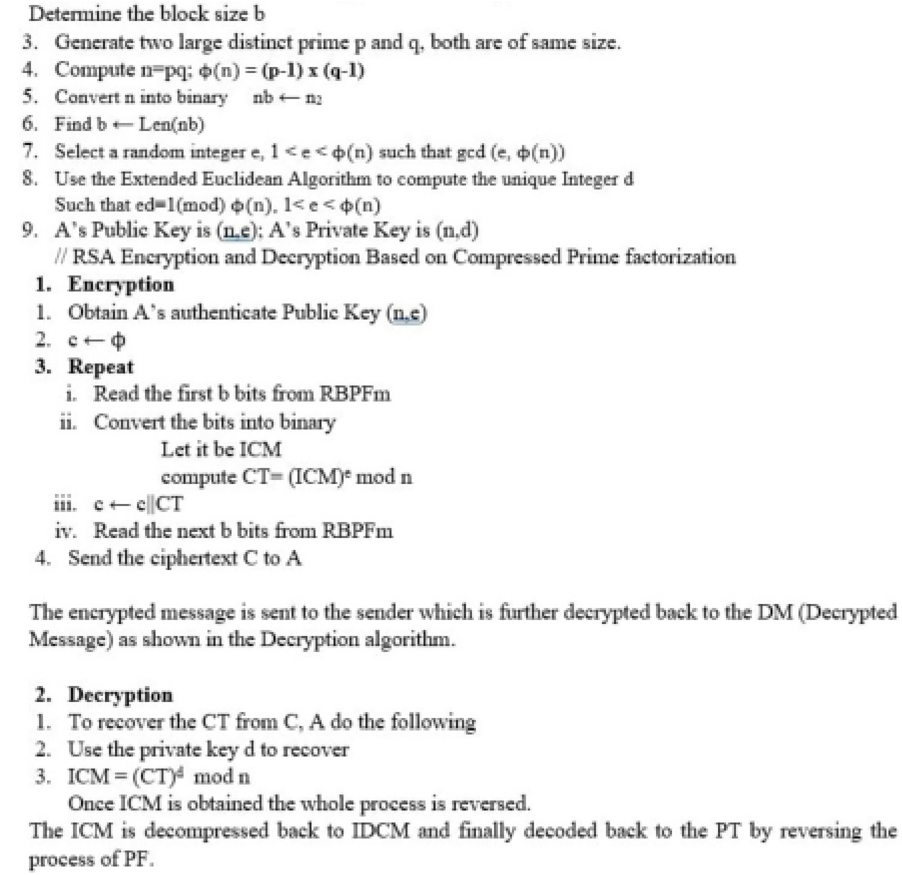
4.3. Dynamic RSA for Encryption and Decryption. The ICM obtained after RLE is taken as the input for converting the ICM into the cipher text which is carried out using Dynamic RSA algorithm. The conventional RSA algorithm usually uses the public key of size 1024 bits or 2048 bits but in this proposed methodology some changes are made, such as limiting the ICM block size to n, which is the product of two powerful prime numbers, p and q, implying that n = p x q. By using the concept of dynamic RSA, the resulted ICM is encrypted as shown in the algorithm 7.

Algorithm 6 RLE Compression



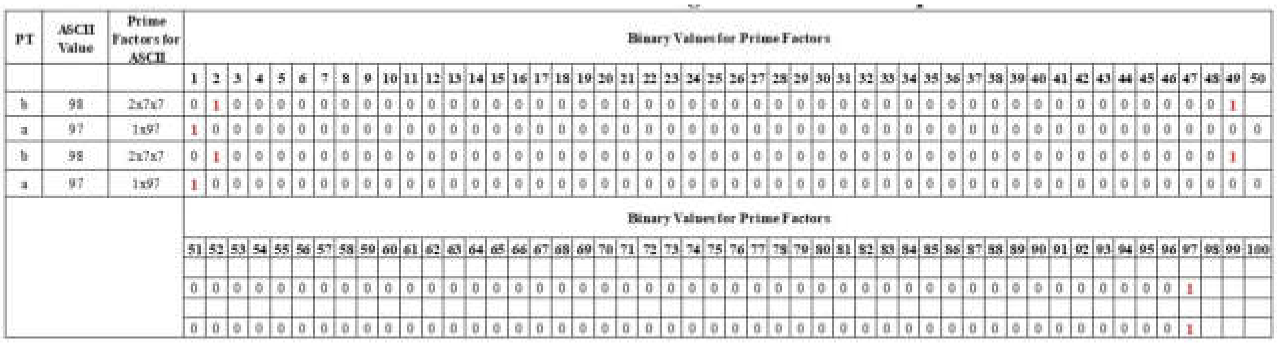
```
RBPFm is Compressed RLE Code for BPFm
Where m = mi, i=1,2,3, l(m)
Input: BPF(m)
Output: RBPFm
1. RBPFm ← $\phi$
2. For each DBPFmi, mi∈m, i=1,2,3…..,n
   If LEN(DBPFmi > 0)
         LC =1
         v = DBPFmi[0]
         v1= DBPFmi[i]
   If (v1 = = v)
         LC=LC+1
   Else
         RBPFm = RBPFm +LC +v
         LC =1
         v= v1
         RBPFm = RBPFm + v1
return RBPFm
```

Algorithm 7 Dynamic RSA for encryption and decryption



```
 Determine the block size b
 3. Generate two large distinct prime p and q, both are of same size.
 4. Compute n=pq; $\phi(n)$ = (p-1) x (q-1)
 5. Convert n into binary    nb ← $n_2$
 6. Find b ← Len(nb)
 7. Select a random integer e, $1 < e < \phi(n)$ such that gcd (e, $\phi(n)$)
 8. Use the Extended Euclidean Algorithm to compute the unique Integer d
    Such that ed≡1(mod) $\phi(n)$, $1< e < \phi(n)$
 9. A's Public Key is (n,e); A's Private Key is (n,d)
    // RSA Encryption and Decryption Based on Compressed Prime factorization
  1. Encryption
  1. Obtain A's authenticate Public Key (n,e)
  2. c ← $\phi$
  3. Repeat
     i. Read the first b bits from RBPFm
    ii. Convert the bits into binary
              Let it be ICM
              compute CT= $(ICM)^e$ mod n
   iii. c ← c||CT
    iv. Read the next b bits from RBPFm
  4. Send the ciphertext C to A

 The encrypted message is sent to the sender which is further decrypted back to the DM (Decrypted
 Message) as shown in the Decryption algorithm.

  2. Decryption
  1. To recover the CT from C, A do the following
  2. Use the private key d to recover
  3. ICM = $(CT)^d$ mod n
     Once ICM is obtained the whole process is reversed.
 The ICM is decompressed back to IDCM and finally decoded back to the PT by reversing the
 process of PF.
```

Table 5.1: Prime factorization encoding to obtain encoded plain text



PT	ASCII Value	Prime Factors for ASCII	Binary Values for Prime Factors																																																	
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
b	98	2x7x7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
a	97	1x97	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	98	2x7x7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
a	97	1x97	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
			Binary Values for Prime Factors																																																	
			51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			

5. Proposed Methodology – Example. The proposed methodology is explained in brief for the plain text ‘baba’ further used as PT. Here the PT is first converted to its equivalent ASCII values which is further processed through all the different levels of security to find the cipher text. Firstly, the EPT is obtained by finding the prime factors for each letter, such as for the letter ‘b’ the equivalent ASCII value is 98 and its equivalent prime factors are 2 x 7 x 7. After calculating the prime factors, a binary sequence of characters are found for this prime factor by giving the magnitude as 1 for all the values of the obtained prime factors and 0 for all the other numbers i.e. for all non-prime factors, and hence the same process is repeated for all the characters of the plain text ‘baba’ as shown in the Table 5.1.

This sequence of characters is moved to process through RLE Compression Algorithm which results in the following output. RLE o p: ICM: 14802195010148021950l. Its Binary equivalent is 1100110101101011111101011 111001001010111011000100101001101101101 which is called as ICM. The ICM is finally converted into cipher text by using the Dynamic RSA for which the value of two large prime numbers p = 57548534591 and the value of q = 57548534663, then calculating the value of n = 3311833837715018027833, result of multiplication of two prime numbers n = p × q whose binary equivalent is 1011001110001000111011010110000010001101110110100 10101010001001100111001.

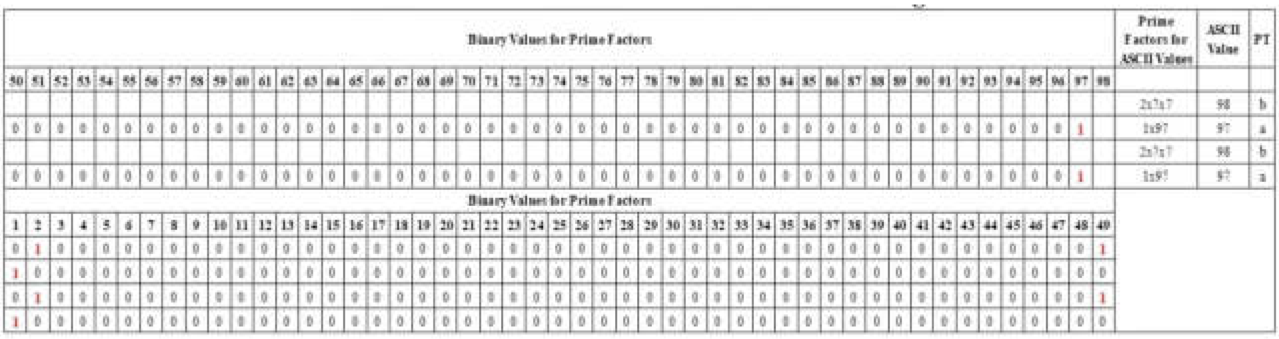
Further $\Phi(n)$ = 57548534590 × 57548534662 = 3311833837599920958580.

Let e = 3311857475573345 67, using the Extended Euclidean algorithm d is calculated as 20173444791987593 48237. Length of number of bits ‘n’ is 72 and hence the block size should be less than 72 wherein here the block size as 64 bits in intermediate message ICM which is the result of RLE Compression Algorithm. Now using the value of ‘e’ and ‘d’ the PT is encrypted as CT, calculated as CT = $(ICM)^e$ mod n = 1480219502148021950133118574755733 4567 mod 3311833837715018027833, resulting in CT= 238379815787132 0965090. To further decrypt the encrypted PT, ICM = CT^d mod n = 2383798157871320965090201734447919875 9348237 mod 3311833837715018027833, resulting in ICM = 14802195021480219501. The final PT is received by the receiver by completely reversing the process, i.e., the CT message is decrypted and decompressed into IDCM.

The obtained IDCM is a sequence of binary numbers, for example for the letter ‘b’ the sequence of binary numbers is 01000000000000000000000000000000000000000000000001 as shown in Table 5.2. Decoding this sequence of bits is done by replacing the binary numbers with respective magnitude values wherever the bit value is ‘1’ which results into the respective prime factors obtained for the ASCII value of b, and finally converting the respective ASCII values back to PT.

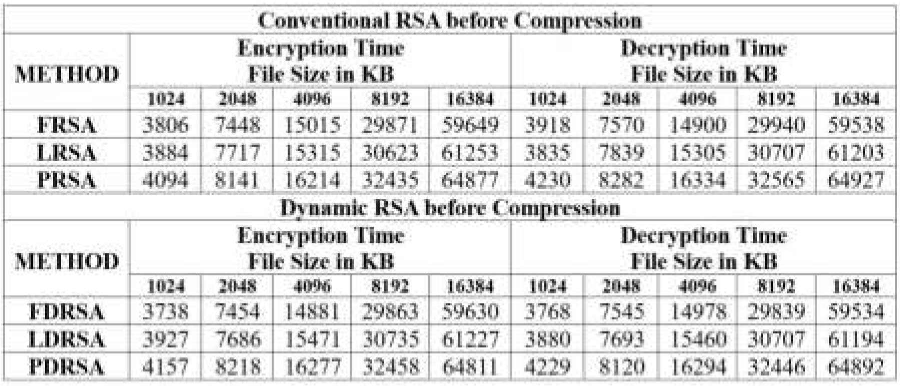
6. Experimental Results. As explained in the proposed methodology the data security is enhanced by encoding the PT using mathematical concept of PF, followed by compression, the data is compressed using RLE and finally the result of compression is encrypted using Dynamic RSA algorithm. This methodology is implemented in VC++ using Core i5 processor for different text files of different sizes. One text file is created which is containing the information of i2k2 Website and the same is used to generate the text files of different sizes such as 1024 KB, 2048 KB , 4096 KB and so on. The security level is calculated for all these files of

Table 5.2: IDCM for prime factorization decoding



Binary Values for Prime Factors																																																	Prime Factors for ASCII Values	ASCII Value	PT
50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98			
																																																	2x7x7	98	b
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		1x97	97	a
																																																	2x7x7	98	b
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		1x97	97	a
Binary Values for Prime Factors																																																			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49			
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			

Table 6.1: Encryption and decryption time for Conventional RSA and Dymamic RSA before compression



Conventional RSA before Compression										
METHOD	Encryption Time File Size in KB					Decryption Time File Size in KB				
	1024	2048	4096	8192	16384	1024	2048	4096	8192	16384
FRSA	3806	7448	15015	29871	59649	3918	7570	14900	29940	59538
LRSA	3884	7717	15315	30623	61253	3835	7839	15305	30707	61203
PRSA	4094	8141	16214	32435	64877	4230	8282	16334	32565	64927
Dynamic RSA before Compression										
METHOD	Encryption Time File Size in KB					Decryption Time File Size in KB				
	1024	2048	4096	8192	16384	1024	2048	4096	8192	16384
FDRSA	3738	7454	14881	29863	59630	3768	7545	14978	29839	59534
LDRSA	3927	7686	15471	30735	61227	3880	7693	15460	30707	61194
PDRSA	4157	8218	16277	32458	64811	4229	8120	16294	32446	64892

different sizes using IBM CAT, which provides a graphical interface for searching, displaying, and analysing data extracted from various cryptographic components. A comparison study is also conducted between the existing method and the proposed method, as well as with conventional and Dynamic RSA with and without the use of a compression algorithm. The different parameters such as encryption time, decryption time, security level and compression ratio are calculated for all file of various sizes which is as shown in the following tables. All these different parameters are calculated using for different file sizes of the cloud data, wherein here the i2k2 cloud desktop as a service is used and the Common Gateway Interface CGI is built for the same.

Table 6.1 and their corresponding graphical representations are shown in Fig. **??** which shows the difference in the encryption and decryption time for text of cloud of different file sizes for all the existing and proposed methodologies using both Prime Factorization Rivest Shamir Adleman PRSA and Prime Factorization Dynamic Rivest Shamir Adleman PDRSA before applying RLE.

The encryption and decryption time taken for the proposed methods PRSA and PDRSA is more than that of the existing methods Fibonacci Rivest Shamir Adleman FRSA and Fibonacci Dynamic Rivest Shamir Adleman FDRSA and Lucas Rivest Shamir Adleman LRSA and Lucas Dynamic Rivest Shamir Adleman LDRSA. The encoding process using PF takes multiple steps like converting the letter to their ASCII values and then finding the prime factors for obtained ASCII value and lastly to encode to their subsequent binary sequence and hence the process of encoding using PF takes more time for encryption and decryption.

Table 6.2 is depicted in Fig. 6.2 which shows the encryption and decryption time for the proposed Prime Factorization Rivest Shamir Adleman Run Length Encoding PRSAR and Prime Factorization Dynamic Rivest Shamir Adleman Run Length Encoding PDRSAR along with existing methods Fibonacci Rivest Shamir Adleman Run Length Encoding FRSAR, Lucas Rivest Shamir Adleman Run Length Encoding LRSAR and Fi-

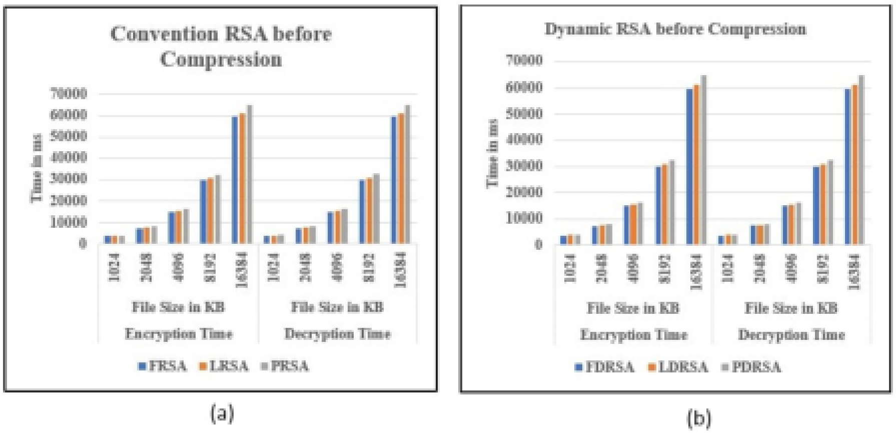
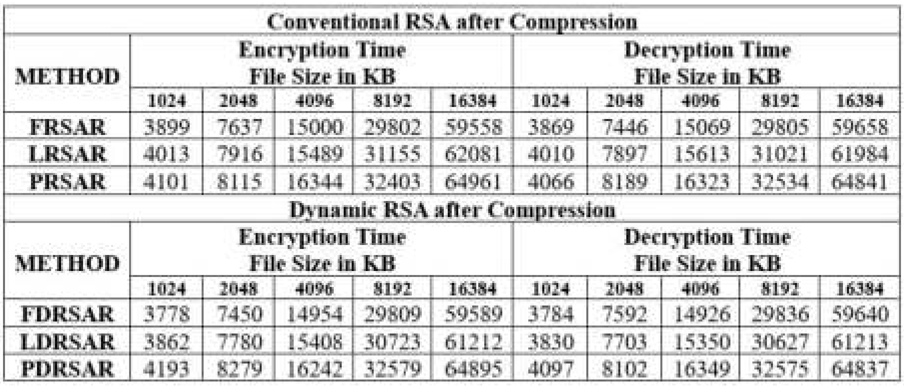


Fig. 6.1: (a) Encryption and decryption time for Conventional RSA before compression; (b) Encryption and decryption time for Dynamic RSA before compression

Table 6.2: Encryption and decryption time for Conventional RSA and Dymamic RSA after compression



Conventional RSA after Compression										
METHOD	Encryption Time File Size in KB					Decryption Time File Size in KB				
	1024	2048	4096	8192	16384	1024	2048	4096	8192	16384
FRSAR	3899	7637	15000	29802	59558	3869	7446	15069	29805	59658
LRSAR	4013	7916	15489	31155	62081	4010	7897	15613	31021	61984
PRSAR	4101	8115	16344	32403	64961	4066	8189	16323	32534	64841
Dynamic RSA after Compression										
METHOD	Encryption Time File Size in KB					Decryption Time File Size in KB				
	1024	2048	4096	8192	16384	1024	2048	4096	8192	16384
FDRSAR	3778	7450	14954	29809	59589	3784	7592	14926	29836	59640
LDRSAR	3862	7780	15408	30723	61212	3830	7703	15350	30627	61213
PDRSAR	4193	8279	16242	32579	64895	4097	8102	16349	32575	64837

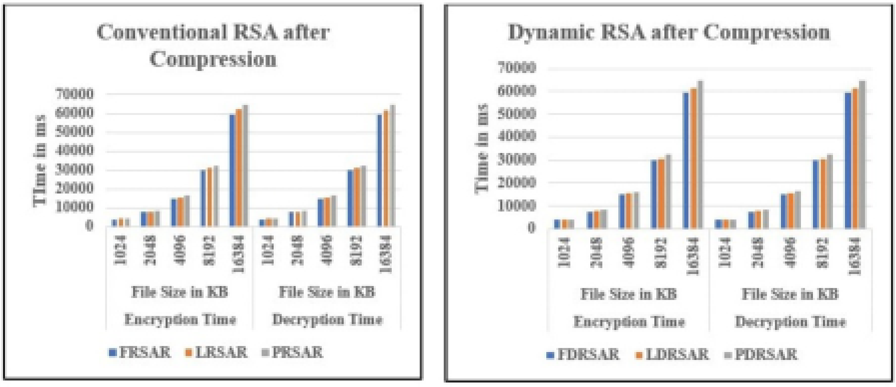
bonacci Dynamic Rivest Shamir Adleman Run Length Encoding FDRSAR, Lucas Dynamic Rivest Shamir Adleman Run Length Encoding LDRSAR after applying RLE Compression algorithm.

The proposed methodology exhibits lesser encryption and decryption time as there is an increase in file size after using RLE compression algorithm in both conventional and dynamic RSA. The encryption time for 16MB file for FRSA is 59649 ms and for FDRSA is 59630 ms as shown in Table 6.1 but the results of Table 6.2 analyse that there is decrease in encryption and decryption time after the addition of the RLE algorithm.

Fig. 6.3 represents the contents of Table 6.3. in which the compression ratio is calculated for proposed methods and PDRSAR and further the comparison is made with existing methods of FRSAR, LRSAR and FDRSAR and LDRSAR. The compression ratio for the above is calculated using the formula as

$$Compression\ ratio = \frac{Uncompressed\ file\ size}{Compressed\ file\ size} \tag{6.1}$$

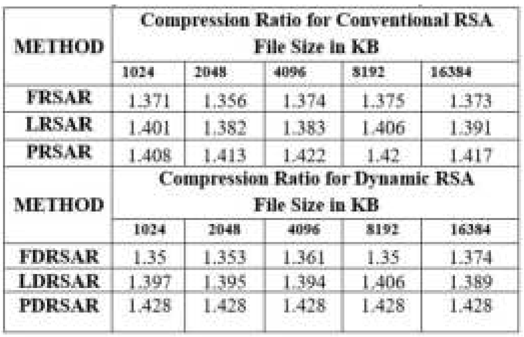
As mentioned in the equation 1 the compression ratio is calculated for different file size and for all existing and the proposed methods and the same is shown in the Table 6.3. The compression ratio for the file size of 1MB for FRSAR is 1024/747 = 1.371, for LRSAR is 1024/731 = 1.401 and that of for the proposed PRSAR method is 1024/727 = 1.408. Similarly, the compression ratio for FDRSAR is 1024/758=1.35, for LDRSAR is 1024/733=1.397 and finally the compression ratio for the proposed methodology PDRSAR is 1024/717=1.428.



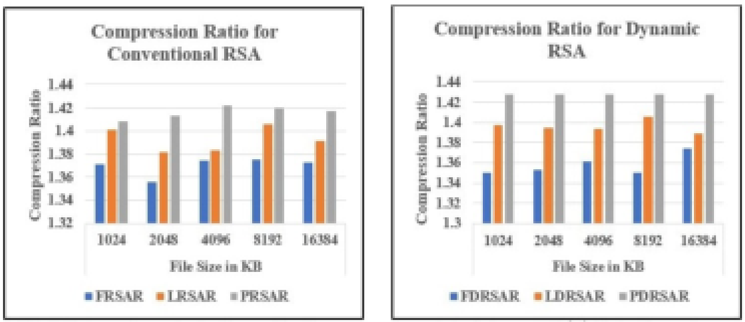
(a) (b)

Fig. 6.2: (a) Encryption and decryption time for Conventional RSA after compression; (b) Encryption and decryption time for Dynamic RSA after compression

Table 6.3: Compression rate for Conventional RSA and Dynamic RSA after compression



METHOD	Compression Ratio for Conventional RSA File Size in KB				
	1024	2048	4096	8192	16384
FRSAR	1.371	1.356	1.374	1.375	1.373
LRSAR	1.401	1.382	1.383	1.406	1.391
PRSAR	1.408	1.413	1.422	1.42	1.417
METHOD	Compression Ratio for Dynamic RSA File Size in KB				
	1024	2048	4096	8192	16384
FDRSAR	1.35	1.353	1.361	1.35	1.374
LDRSAR	1.397	1.395	1.394	1.406	1.389
PDRSAR	1.428	1.428	1.428	1.428	1.428



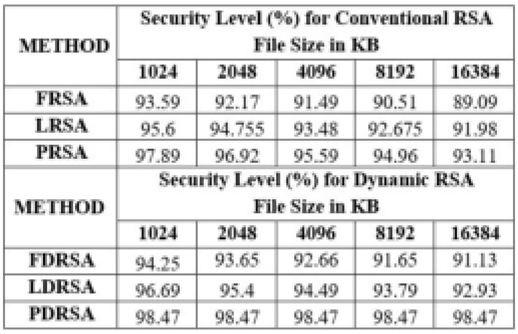
(a) (b)

Fig. 6.3: (a) Compression for Conventional RSA after compression; (b) Compression ratio for Dynamic RSA after compression

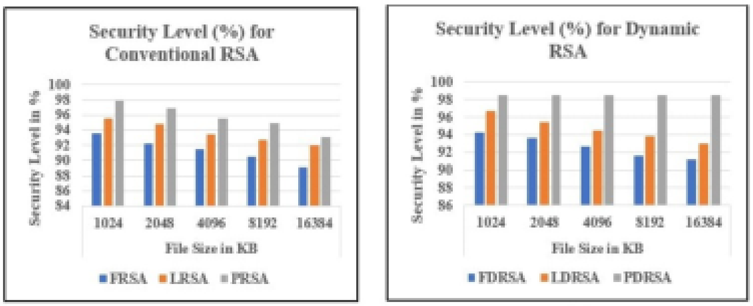
This analysis clears that the compression ratio is comparatively better for the proposed methods PRSAR and PDRSAR than that of the existing methods.

The security level is analysed for the proposed and existing methods and the comparison analysis is made

Table 6.4: Security level for Conventional RSA and Dynamic RSA before compression



METHOD	Security Level (%) for Conventional RSA File Size in KB				
	1024	2048	4096	8192	16384
FRSA	93.59	92.17	91.49	90.51	89.09
LRSA	95.6	94.755	93.48	92.675	91.98
PRSA	97.89	96.92	95.59	94.96	93.11
METHOD	Security Level (%) for Dynamic RSA File Size in KB				
	1024	2048	4096	8192	16384
FDRSA	94.25	93.65	92.66	91.65	91.13
LDRSA	96.69	95.4	94.49	93.79	92.93
PDRSA	98.47	98.47	98.47	98.47	98.47

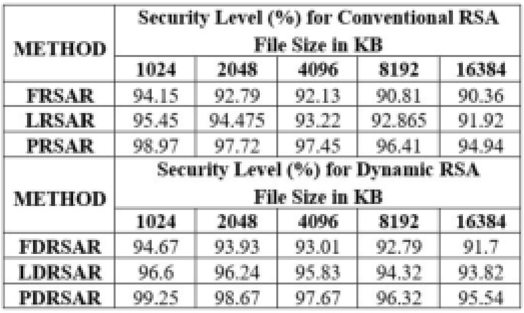


(a)

(b)

Fig. 6.4: (a) Security level (%) for Conventional RSA before compression; (b) Security level (%) for Dynamic RSA before compression

Table 6.5: Security level for Conventional RSA and Dynamic RSA after compression



METHOD	Security Level (%) for Conventional RSA File Size in KB				
	1024	2048	4096	8192	16384
FRSAR	94.15	92.79	92.13	90.81	90.36
LRSAR	95.45	94.475	93.22	92.865	91.92
PRSAR	98.97	97.72	97.45	96.41	94.94
METHOD	Security Level (%) for Dynamic RSA File Size in KB				
	1024	2048	4096	8192	16384
FDRSAR	94.67	93.93	93.01	92.79	91.7
LDRSAR	96.6	96.24	95.83	94.32	93.82
PDRSAR	99.25	98.67	97.67	96.32	95.54

before using compression and after using compression algorithm.

The results of Table 6.4 shows that the security level is improved in both cases of conventional and dynamic RSA for the proposed methods of PRSA and PDRSA compared to that of existing methods of FRSA, FDRSA and LRSA, LDRSA and the same is graphically represented in Fig. 6.4. The results analysis also depicts that there is a decrease in security level as the size of the file increases for the conventional RSA but that of the proposed methodology of PDRSA the security level remains same for the varying file sizes.

Similarly, the results of Table 6.5. shows the security level which is again compared for the proposed and all the existing methods after applying the RLE compression algorithm which intern is graphically represented in

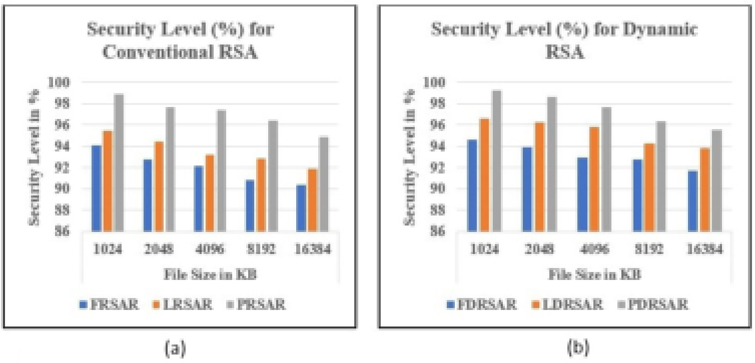


Fig. 6.5: (a) Security level (%) for Conventional RSA after compression; (b) Security level (%) for Dynamic RSA after compression

Fig. 6.5. The security level is proved to be improved for the proposed PRSAR and PDRSAR compared to that of the existing LRSAR, LDRSAR and FRSAR, FDRSAR. The encryption and decryption time is compromised for the proposed methodology as the major aim was to achieve the security level which intern is proved by achieving the results of 99.25%for 1MB file obtained finding multiple intermediate messages.

The encryption time and decryption time is always more when process the plain text through the compressed cryptosystems along with encoding i.e., as in here the plain text is first converted to M1, then moved for compression and lastly it is encrypted. But during this entire process the security level is increased compared to the conventional RSA algorithm applied without any intermediate message like M1 and M2.

Even though the encryption time and decryption time are more for PF compared to Fibonacci and Lucas, the results showed a good progress in the security level and also in the compression ratio. The security level of the plain text is more for the proposed Dynamic RSA Prime factorization compared to the existing Dynamic RSA Fibonacci and Dynamic RSA Lucas which is as shown in the Table 6.5. Along with increase in the security level the concentration is also given for the compression ratio which found to be more efficient for the proposed methodology.

7. Conclusion. The proposed method could be applied for the text data of any cloud-based application and the implementation of the prime factorization to enhance the security level of the cloud data proved to be 95.54% for 16 MB files for the proposed PDRSAR which is improved than LDRSAR that resulted with security level of 93.82% and that for FDRSAR achieving the results of 91.7%. The compression ratio for the proposed method proved to be 1.428 which is improved than that of the existing methods of LDRSAR with 1.389 and that for FDRSAR with 1.374. The compression algorithm helped to improve the compression ratio and also helped to increase the security level by developing the intermediate messages at each stage. The complete work is implemented for the text of the cloud and hence in future it is assured to work the process for the images and then with that of the combined approach wherein in all the three approaches the concentration will be on enhancing the security for the cloud data irrespective of the type of data.

Acknowledgement. I extend my sincere gratitude to REVA University for their invaluable support and guidance throughout the course of this research, which greatly facilitated its successful completion.

REFERENCES

[1] Mohammed Aamir Ali, Muhammad Ajmal Azad, Mario Parreno Centeno, Feng Hao, Aad van Moorsel, “Consumer-facing technology fraud: Economics, attack methods and potential solutions”, Future Generation Computer Systems,Volume 100,2019, Pages 408-427,ISSN 0167-739X.

[2] V. Pavani, P. S. Krishna, A. P. Gopi, and V. L. Narayana, “Secure data storage and accessing in cloud computing using enhanced group based cryptography mechanism,” in Materials Today: Proceedings, Dec. 2020, pp. 1-5, doi: 10.1016/j.matpr.2020.10.262.

[3] A. Devi and K. Mani, “CSEIT1831152 | Enhancing Security in RSA Cryptosystem Using Burrows-Wheeler Transformation and Run Length Encoding,” 2018. [Online]. Available: www.ijsrcseit.com.

[4] Luluk Anjar Fitriya, Tito Waluyo Purboyo, Anggunmeka Luhur Prasasti, “A Review of Data Compression Techniques”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 19 (2017) pp. 8956-8963.

[5] I. Sandhya Rani and Bondu Venkateswarlu, “A Systematic Review of Different Data Compression Technique of Cloud Big Sensing Data” ICCNCT 2019, LNDECT 44, pp. 222–228, 2020. doi.org/10.1007/978-3-030-37051-0-25

[6] K. Mani and A. Devi, “Enhancing security in cryptographic algorithm based on LECCRS,” 2017. [Online]. Available: http://en.wikipedia.org/wiki/

[7] M. Tajammul and R. Parveen, “Auto encryption algorithm for uploading data on cloud storage,” International Journal of Information Technology (Singapore), vol. 12, no. 3, pp. 831–837, Sep. 2020, doi: 10.1007/s41870-020-00441-9.

[8] P. William, A. Choubey, G. S. Chhabra, R. Bhattacharya, K. Vengatesan, and S. Choubey, “Assessment of Hybrid Cryptographic Algorithm for Secure Sharing of Textual and Pictorial Content,” in Proceedings of the International Conference on Electronics and Renewable Systems, ICEARS 2022, 2022, pp. 918–922. doi: 10.1109/ICEARS53579.2022.9751932.

[9] K. Mani and A. Devi, “Modified DES using Different Keystreams Based On Primitive Pythagorean Triples,” International Journal of Mathematical Sciences and Computing, vol. 3, no. 1, pp. 38–48, Jan. 2017, doi: 10.5815/ijmsc.2017.01.04.

[10] Amruta Gadad, Devi Anbusezhiyan, “Cloud security: literature survey”, International Journal of Electrical and Computer Engineering (IJECE), Vol. 13, No. 4, August 2023, pp. 4734 4742, ISSN: 2088-8708, DOI: 10.11591/ijece.v13i4.pp4734-4742

[11] J. Zalaket and J. Hajj-Boutros, “Prime factorization using square root approximation,” Computers and Mathematics with Applications, vol. 61, no. 9, pp. 2463–2467, May 2011, doi: 10.1016/j.camwa.2011.02.027.

[12] R. Jakimczuk, “Sums of Prime Factors in the Prime Factorization of Smooth Numbers Diophantine equations View project Prime Numbers View project Sums of Prime Factors in the Prime Factorization of Smooth Numbers”.

[13] Wid Akeel Awadh, Ali Salah Alasady, Mohammed S. Hashim, “A multilayer model to enhance data security in cloud computing”, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 32, No. 2, November 2023, pp. 1105 1114, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v32.i2.pp1105-1114.

[14] Sunday Adeola Ajagbe, Oluwashola David Adeniji, Adedayo Amos Olayiwola, Seun Femi Abiona, “Advanced Encryption Standard (AES)-Based Text Encryption for Near Field Communication (NFC) Using Hufman Compression”, SN Computer Science (2024) 5:156, https://doi.org/10.1007/s42979-023-02486-6.

[15] Shiladitya Bhattacharjee, Himanshi Sharma, Tanupriya Choudhury, Ahmed M. Abdelmoniem, “Leveraging chaos for enhancing encryption and compression in large cloud data transfers”, The Journal of Supercomputing.

[16] N. Sugirtham, R. Sherine Jenny, B. Thiyaneswaran, S. Kumarganesh, C. Venkatesan, K. Martin Sagayam, Lam Dang, Linh Dinh, Hien Dang, “Modifed Playfair for Text File Encryption and Meticulous Decryption with Arbitrary Fillers by Septenary Quadrate Pattern”, International Journal of Networked and Distributed Computing, https://doi.org/10.1007/s44227-023-00019-4.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jun 8, 2024
Accepted: Aug 29, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1104–1115, DOI 10.12694/scpe.v26i3.4254

ENHANCED EARLY DIAGNOSIS OF LIVER DISEASES USING FEATURE SELECTION AND MACHINE LEARNING TECHNIQUES ON THE INDIAN LIVER PATIENT DATASET

ARUN GANJI [*], D. USHA [†] AND P.S. RAJAKUMAR [‡]

Abstract. Liver diseases are a significant global health concern, with timely diagnosis crucial for effective treatment and prevention of further damage. This study addresses the challenge of early liver disease detection using machine learning techniques applied to the Indian Liver Patient Dataset (ILPD). Our proposed method comprises a four-phase approach: (1) initial model training using five machine learning algorithms - Multilayer Perceptron (MLP), Support Vector Machine (SVM), Decision Trees (DT-CART), Light Gradient Boosting Machine (LGBM), and Logistic Regression (LR) - on the original dataset; (2) feature selection using Forward Selection (FS) to identify the most relevant attributes; (3) model retraining with the selected features; and (4) model optimization to enhance prediction accuracy. The dataset was split into 80% training and 20% testing sets, with 10-fold cross-validation applied throughout. Our findings demonstrate the significant impact of feature selection and model optimization on algorithm performance. The Light Gradient Boosting Machine (LGBM) emerged as the top-performing model, achieving an accuracy of 82.12% after optimization, compared to its initial 76.21%. LGBM also showed balanced performance across specificity, sensitivity, precision, and F1-score metrics. This study contributes to the field by presenting a comprehensive approach to liver disease prediction, emphasizing the importance of feature selection and model optimization in improving diagnostic accuracy.

Key words: Indian Liver Patient Dataset (ILPD), Machine learning (ML), Forward Selection (FL), Classification Algorithms.

1. Introduction.

The liver, the largest internal organ in the human body, plays a pivotal role in numerous physiological processes, performing more functions than any other organ. Its significance in maintaining overall health cannot be overstated, making the early detection and treatment of liver diseases a critical medical priority. Liver enzymes, particularly Aspartate Aminotransferase (SGOT) and Alanine Aminotransferase (SGPT), are crucial indicators in diagnosing liver diseases. Various factors, including lifestyle habits such as smoking and alcohol consumption, can trigger liver diseases and elevate these enzyme levels [1]. Moreover, conditions like diabetes, hepatitis B, and hepatitis C can lead to liver damage and, if left untreated, progress to liver failure. The consequences of severe liver damage are dire, often necessitating organ transplantation or resulting in mortality [2]. These facts underscore the urgent need for early and accurate diagnosis of liver diseases.

In recent years, the rapid advancement of technology has ushered in a new era of medical diagnostics, with machine learning emerging as a powerful tool in various healthcare domains. Machine learning methods can analyze vast amounts of data, both structured and unstructured, to create predictive models using statistical and mathematical techniques. The efficacy of these predictions is intrinsically linked to the quality of the underlying model.

In the realm of classification algorithms, the selection of relevant features plays a crucial role in model performance. Not all attributes in a dataset contribute equally to the model's predictive power. To enhance model accuracy and efficiency, researchers often employ feature selection methods to identify the most relevant attributes. This process is vital in creating effective machine learning classification models for medical diagnostics.

The present study addresses the critical need for improved liver disease diagnosis by proposing a hybrid model that combines feature selection with advanced machine learning techniques. We utilize the Indian Liver Patient Dataset (ILPD) and employ the forward selection method, a wrapper approach, for feature selection. The study then applies various machine learning algorithms, including Support Vector Machine

[*] Dr. M.G.R. Educational and Research Institute, Chennai-95, Tamilnadu, India (arun.ganji@gmail.com).
[†] Dr. M.G.R. Educational and Research Institute, Chennai-95, Tamilnadu, India (usha.cse@drmgredu.ac.in).
[‡] Dr. M.G.R. Educational and Research Institute, Chennai-95, Tamilnadu, India (rajakumar.subramanian@drmgredu.ac.in).

(SVM), Multilayer Perceptron (MLP), Decision Trees (DT), Logistic Regression (LR), and Light Gradient Boosting Machine (LGBM), to diagnose liver failure using the most important features identified through feature selection. Furthermore, we enhance the diagnostic accuracy by optimizing the hyperparameters of these machine learning models.

Our approach comprises four key stages:

1. Initial application of classification algorithms on the complete dataset.
2. Feature selection to create a subset of important attributes, followed by the application of classification algorithms on this subset.
3. Model improvement through hyperparameter optimization and comparison of prediction accuracies across the first three stages.
4. Comparative analysis of our optimized models' performance with existing literature using the same dataset.

This comprehensive approach aims to significantly improve the accuracy of liver disease prediction, potentially leading to earlier diagnoses and more effective treatment strategies. By combining feature selection, advanced machine learning techniques, and model optimization, we address the pressing need for more accurate and efficient diagnostic tools in hepatology.

The subsequent sections of this paper provide a detailed review of related studies, a comprehensive description of our methodology, including the dataset and machine learning methods employed, and a thorough comparison of our results with existing literature. Through this research, we aim to contribute to the ongoing efforts to enhance liver disease diagnosis and, ultimately, improve patient outcomes.

2. Related Study. Numerous studies have been carried out on various datasets related to liver diseases. This section provides a summary of some of the key studies that are prominent in the literature on this topic. Literature [3] modelled the SVM method in the diagnosis of diabetes and chronic liver disease on Diabetes, BUPA and ILPD datasets in the MATLAB study environment. They achieved 63% success in the diagnosis made with the first 4 attributes of the BUPA data set, 70% with the first 6 attributes, and 70% with the first 8 attributes. On the ILPD data set, they stated that the diagnosis made with the first 4 attributes of the data set was 71%, 73% with the first 6 attributes, and 73.2% with the first 8 attributes.

In the literature [4] Gulia et al. used J48, Random Forest (RF), MLP, SVM, and Bayesian Net machine learning methods towards diagnose liver diseases on the ILPD dataset. In the first stage, they made diagnostic success measurements with the original version of the data set without selecting attributes on the data set. In the first stage, they achieved the highest diagnostic success rate with 71.35% from the SVM classification method. They stated that the other algorithms were successful in RF 70.31%, MLP 68.25%, J48 68.70%, and Bayes Net 67.23%, respectively. In the second stage, they applied the Greedy Stepwise (Greedy algorithm) feature selection on the dataset to identify significant features. They found that the crucial attributes are Total Bilirubin and Direct Bilirubin, Total Proteins, Albumin and A/G ratio. In the second stage, the accuracy of the RF algorithm was determined with a rate of 71.86%. Other methods were successful in SVM 71.34%, J48 70.65%, MLP 70.82%, and Bayeses Net 68.11%, respectively.

Literature [5] used ANN, RF, Functional Tree and Radial Based Functional machine learning methods for the diagnosis of liver failure on BUPA and ILPD datasets. 86.95% of BUPA data were used as training (300 samples) and 13.05% (45 samples) as tests. For the ILPD data set, 87.48% (510 samples) are utilised for training and 12.52% (73 samples) are used for testing. They tested ANN using the MATLAB environment along with other methods in the WEKA tool. In the analyses, they used a 10-fold cross-validation test technique for the methods they used in WEKA and 10 hidden layers and a network of forward-fed neurons for ANN. As a result of this study, the highest diagnostic success was 76% for the BUPA dataset with ANN and 78% for the ILPD dataset.

In the research [6] author evaluated the ILPD dataset in two different stages. They applied a feature model also comparative study to enhance forecast accuracy. In the first stage, they applied a minimum maximum normalization filter to the original data set. In the second stage, they identified the attributes containing important features by using PSO (Particle Swarm Optimization) feature selection on data set. They identified the important attributes were Direct Bilirubin, Total Bilirubin, Total Proteins, A/G ratio, Albumin, SGOT, SGPT, and Alkphos. In the first stage, Bayes Net was the most successful method with an accuracy rate of

74.25%. In other methods, J.48 achieved 73.32%, MLP 69.22%, SVM 71.44%, and RF 68.43%, respectively. In the second stage, the J48 method was the most successful algorithm with a rate of 95.04%. In the relevant study, Bayes' Net was 90.33%, RF was 80.22%, MLP was 77.54% and SVM was 73.44%.

Alice [7] applied DT, RF, Naive Bayes, AMM and SVM machine learning methods to identify liver failure using ILPD dataset. R programming language was used in the study. The DT algorithm achieved the highest diagnostic success rate with 81%. Among the additional methods, RF was 77%, ANN was 71%, SVM was 77%, and Naive Bayes was 37%.

Literature [8] examined the success rates of machine learning methods in medical datasets. The designated medical datasets are Breast Cancer Data, Cryotherapy, Chronic Kidney Disease, Hepatitis, ILPD and Immunotherapy. They analyzed the datasets with Naive Bayes, J48, MLP, JRip, IBk and Bagging machine learning methods. In their study, they used 10-fold cross-validation. In the ILPD data set, they obtained the highest diagnostic success rate from the Bagging method with a rate of 69.30%. In other methods, they stated that they achieved 68.95% success in MLP, 68.78% in J48, 66.38% in JRIP, 64.49% in IBk and 55.75% in Native Bayes, respectively.

In the literature [9], the author focused on predicting liver disease built on a software engineering methodology using the trait selection and classification technique. Intelligent liver disease prediction software (ILDPS) was developed using feature selection and classification prediction techniques based on a software engineering model. They applied LR, RF, SMO, Naive Bayes, IBk and J48 machine learning approaches to find accuracy on the ILPD dataset. They evaluated the dataset in two stages. In the first stage, they made diagnostic success measurements with the original version of the data set without selecting attributes upon data set. In second stage, they utilized the Greedy Stepwise feature selection algorithm to identify significant features within the dataset. They determined that the key attributes are Total Bilirubin, Alkphos, Direct Bilirubin, SGOT, and SGPT. Their accuracy rate was tested using 10x cross-validation. In the first stage, the highest diagnostic success rate of 72.53% was achieved using the RF machine learning method. Among the other methods, LR was 72.50%, SMO was 71.35%, J48 was 68.78%, IBk was 64.15% and Naive Bayes was 55.74%. In the second stage, the LR machine learning method achieved the highest diagnostic success rate at 73.36%. In further methods, RF was 71.87%, SMO was 71.36%, J48 was 70.67%, IBk was 67.41% and Naive Bayes was 55.90%.

In the research [10] author used LR, SVM, K-Nearest Neighbour (K-NN) and on ILDP datasets, ANN machine learning approaches used to diagnose liver failure. They presented the model with the maximum accuracy as a Graphical User Interface (GUI) using the Tkinter package in Python. The utmost diagnostic success rate has been obtained from the ANN machine learning technique by a rate of 92.80%. Of the other methods, SVM was 75.04%, LR was 73.23% and C-NN was 72.05%. In ANN, number of inputs is set to 10, number of hidden layers is set to 2, The first hidden layer is configured with 400 neurons, and the second hidden layer is also configured with 400 neurons.

This study we developed and presented a hybrid model and a comparative analysis to improve the prediction accuracy of liver failure patients in four phases.

3. Materials and Methods. In this study, the ILPD dataset shared for research in the UCI machine learning pool has been used for liver failure disease diagnosis. There are 583 specimens in the ILPD dataset. The data were split into 80% training set (466 samples) and 20% test set (117 samples). A random state value of 42 was used to use the same training and test set in each data set separation process. Gender attribute was not taken into account in the data set. This study presents a hybrid model aims to enhance the prediction accuracy for liver failure patients across four phases. In first stage, SVM, MLP, DT, LR and LGBM machine learning methods were applied to the original dataset with the attributes ”Total Bilirubin, Age, Direct Bilirubin, Alanine Aminotransferase, Alkaline Phosphatase, Total Protein, Albumin, Albumin/Globulin Ratio” and diagnostic prediction successes were measured. In the second stage, forward selection, which is one of the Wrapper methods, was used for feature selection on the data set. After the feature selection, the predictive success of the diagnosis of liver failure disease was measured by SVM, MLP, DT, LR and LGBM machine learning methods using the attributes ”Direct Bilirubin, Age, Alanine Aminotransferase, Alkaline Phosphatase”. In the third stage, the models were re-established by applying model improvement to the SVM, MLP, DT, LR and LGBM machine learning methods applied in the second stage, and their accuracy performance was increased. The forecast accuracy performances of the first three stages were compared. In the fourth stage, the accuracy

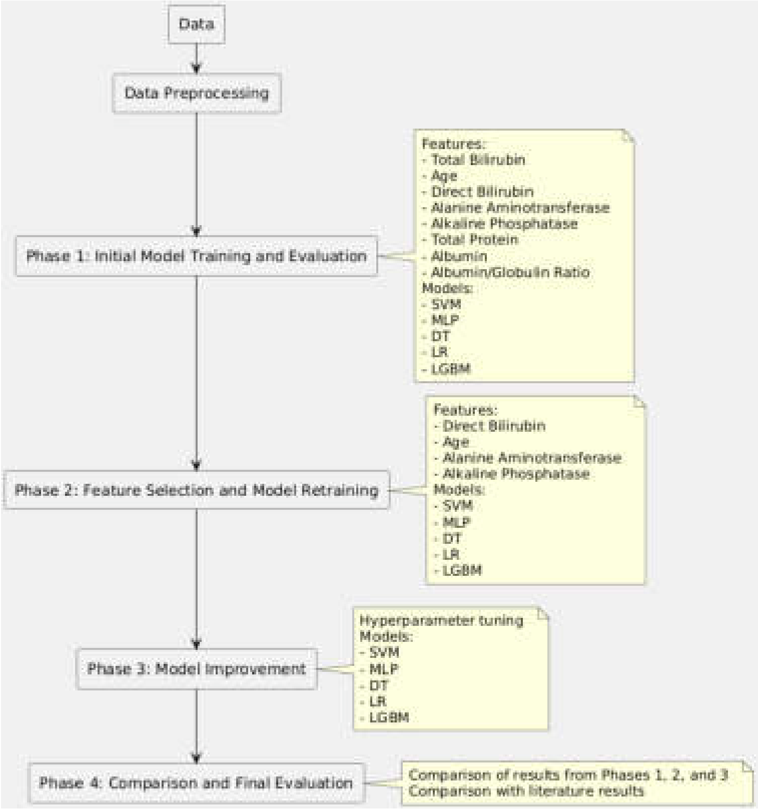


Fig. 3.1: Proposed Architecture.

performance results of the classification methods with model improvement were compared with the data set used in this study in the literature. The modelling of machine learning methods was carried out in the JupyterLab environment by means of the Python programming language. While measuring the success of machine learning methods, cross-validation method (cv) was used as 10 times.

3.1. Proposed Architecture. The proposed architecture Figure 3.1 and the algorithm (Algorithm 1) for liver disease prediction consists of four main phases:

Data Preparation and Initial Model Training: The ILPD dataset is preprocessed, with the gender attribute removed. The data is then split into 80% training (466 samples) and 20% test (117 samples) sets, using a random state of 42 for reproducibility. Five machine learning algorithms (SVM, MLP, DT, LR, and LGBM) are trained on the original dataset using all eight attributes: Total Bilirubin, Age, Direct Bilirubin, Alanine Aminotransferase, Alkaline Phosphatase, Total Protein, Albumin, and Albumin/Globulin Ratio.

Feature Selection: The Forward Selection method, a wrapper approach, is applied to identify the most important features in the dataset. This process reduces the feature set to four key attributes: Direct Bilirubin, Age, Alanine Aminotransferase, and Alkaline Phosphatase. Model Retraining with Selected Features: The five machine learning algorithms are retrained using only the four selected features. This step aims to improve model performance by focusing on the most relevant attributes.

Model Optimization and Evaluation: Each model undergoes hyperparameter tuning to further enhance its performance. The optimized models are then evaluated using 10-fold cross-validation. Performance metrics such as accuracy, specificity, sensitivity, precision, and F1-score are calculated and compared across all phases.

Table 3.1: Dataset properties

Sl#	Properties	Descrption	Data Type	Property Range Value
1	Age	Age	Numeric	[4-90]
2	Gender	Gender	Norminal	”male” or ”female”
4	TB	Total Bilirubin	Numeric	[0.4-75]
5	DB	Direct Bilirubin	Numeric	[0.1-19.7]
6	AlkPhos	Alkalen fosphatase	Numeric	[63-2110]
7	Sggt	Alanine Aminotransferase	Numeric	[10-2000]
8	Sgot	Aspartate Aminotransferase	Numeric	[10-4029]
9	TP	Total Protein	Numeric	[2.7-9.6]
10	ALB	Albumen	Numeric	[0.9-5.5]
11	A/G Ratio	Albumin/Globulin Ratio	Numeric	[0.3-2.8]
12	Selector Field	Selective Domsin Information	Numeric (12)	1-Sick,2-Not Sick

3.2. Data Set.

Dataset Selection. For this study, we chose the Indian Liver Patient Dataset (ILPD) [11]for our analysis and model development. The ILPD was selected for several key reasons:

Relevance and Specificity. The ILPD focuses specifically on liver patients, aligning perfectly with our research goal of improving liver disease diagnosis. It contains crucial liver function tests and patient attributes essential for identifying liver disorders.

Diversity and Representation. This dataset represents a diverse population from India, a country with a high prevalence of liver diseases, enhancing the generalizability of our findings. Dataset Quality: The ILPD is a well-curated dataset collected by reputable medical institutions, ensuring high data quality and reliability.

Balanced Representation. It includes both liver patients and non-liver patients, providing a balanced dataset crucial for developing accurate classification models.

Feature Richness. The ILPD contains various features including age, gender, and multiple blood tests, allowing for comprehensive analysis and feature selection experiments.

Benchmark Status. As a widely used dataset in liver disease prediction research, the ILPD enables direct comparison of our results with other state-of-the-art methods in the literature [16].

Public Availability. Being part of the UCI Machine Learning Repository, the ILPD is freely available, promoting reproducibility and further research in the field.

Challenging Nature. The dataset presents a complex classification problem, making it an excellent testbed for evaluating our proposed hybrid model and feature selection approach [17].

By utilizing the ILPD, we aim to contribute meaningfully to liver disease diagnosis research while ensuring our results are comparable, relevant, and potentially impactful for a significant patient population. This choice aligns well with our research objectives and the broader goal of improving early detection of liver diseases [18].

The total 583 sample patient records in the repository. Of these records, 441 were male and 142 were female. Of the patient records, 416 had liver disease and 167 did not have liver disease. The total men with liver disease is 324 and total women with liver disease is 92. There are 11 features in the dataset. While 10 of them can be used as attributes, the eleventh feature is the field where the presence of the disease is shown. SGPT and SGOT, which are found in attributes, are used under different names today. In the new terminology, SGOT is called ALT, and SGPT is called AST. Detailed information about the dataset characteristics is presented in Table 3.1.

3.2.1. Normalization Filter. The high number of changes between data affects the classification methods learning accuracy in some way. Normalization aims to eliminate discrepancies between mathematical operations and data, facilitating easier data comparison. In this study, the data were normalized using the Standard Scaler from the sklearn.preprocessing library in Python, applied to both MLP and SVM algorithms.

3.2.2. Attribute Selection. In this study, among wrapper methods (SFS) Step Forward Selection is one of them, was used to select features in the data set. In the SFS method, the cycle process is initiated

Table 3.2: Important attributes obtained after the SFS method.

Attribute Name	Abbreviation
Direct Bilirubin	DB
Age	Age
Alanine Aminotransferase	Sgpt
Alkaline Phosphatase	AlkFos

Table 3.3: Hyperparameters used in MLP.

Activation	relu
alpha	0.1
hidden_layer_size	(10,10,10,10,10)
random_state	42
max_iter	215

with the attribute that contributes the most to the performance of the established model, that is, it has the highest correlation with the dependent variable. Other attributes are checked in order according to the specified severity level ( ). Attributes that satisfy the condition based on the level of importance are added to the model. Attributes that don't meet the initial severity level requirement are not added to the model. Once all variables have been checked, the cycle ends. In this study, the attribute was selected with a significance level of 95%. The attributes determined by the SFS method has been presented in Table 3.2.

3.3. Support Vector Machine. Vapnik and his developed the Support Vector Machine algorithm in the 1990s. The Support Vector Machine, which is one of the supervised machine learning algorithms, is used to distinguish between binary base class data by applying statistical processing on it [12].

Since a classification belonging to two classes will be used in our data set, linear support vector classification is preferred. Reducing the value of the maximum number of iterations during the model setup phase while model optimization increased the accuracy value. Therefore, it max_iter the best performance in the accuracy rate, the hyperparameter is obtained as 23. While creating the improved model, the random state value was used as 42 and the maximum number of iters was used as 23. In this study, the Linear SVM machine learning method called LinearSVC, which is available in the Python programming language sklearn.svm library, was used.

3.4. Multilayer Sensors. Multilayer Perceptrons consist of input, intermediate and output layers. Unlike a single-layer sensor, the intermediate layer acts as a bridge between the output layer and the input layer. The middleware evaluates the inputs from the input layer against the problem before sending them to the output layer. As a result of the evaluation, a better decision is made according to the problem. The number of interlayers can be increased according to the condition of the problem [13].

In MLP, a five-layered model consisting of 10 neurons in each hidden layer was used in the improved model studies. The corrected linear unit function is preferred for activation function in hidden layers. A value of 0.1 as the L2 penalty parameter increased the accuracy. By adding the maximum number of iterations to 215, the best performance in accuracy was achieved. The hyperparameters that are changed when building an optimized model are shown in Table 3.3. In this study, the MLPClassifier classification algorithm in the Python programming language sklearn.neural_network library was used.

3.5. Logistic Regression. It is a statistical method that solves problems. It specifies two possible (0 or 1) outcomes by regression analysis of the dataset [13]. The outcome dependent variable used in this study is kept numerically (1-liver patient, 2-not liver disease).

In the improved model studies in LR, it has been observed that the use of liblinear as the preferred algorithm for optimization problems increases the success because the size of our data set is small.

Table 3.4: Hyperparameters used in logistic regression.

random_state	42
multi_class	ovr

Table 3.5: Hyperparameters used for the decision tree.

max_features	auto
max_depth	5
random_state	42

Table 3.6: Hyperparameters used in LGBM.

n_estimators	150
learning_rate	0.2
random_state	42

Since our result variable belongs to 2 classes (1-liver patient, 2-not liver disease), the multi-class feature was used as ovr. The hyperparameters that are changed when creating an optimized model are shown in Table 3.4. In this study, the LogisticRegression classification method in the Python programming language sklearn.linear_model library was used.

3.6. Decision Trees. Algorithms in supervised learning most widely used one is Decision Trees. The primary goal is to convert complex structures in the dataset into simpler ones. With decision trees, both numerical and categorical data can be processed. It is one of the commonly used approaches in classification because the test and training are fast and the results can be interpreted visually easily. It consists of roots, nodes and branching. In decision trees, operations are concluded in two steps. In step one, creation of the tree. In step two, classification rules are created from the created tree. Depending on the algorithms used to create the decision tree, the construction of the decision tree may vary [14]. As given in Algorithm 2, the tree is built using a top-down, recursive approach known as recursive partitioning. At each step, the algorithm chooses the attribute that best splits the set of items, typically using measures like Gini impurity or information gain.

In this study, the Decision Tree Classifier machine learning method in the Python programming language sklearn.tree library was used. The decision tree was created with the Classification and Regression Tree (CART) algorithm. Entropy algorithm was used for branching. Limiting the maximum depth of the tree to 5 increased the accuracy rate. For the best division, the max_features parameter auto corresponding to max_features=sqrt(n_features) was used. The hyperparameters that are changed when building an optimized model are shown in Table 3.5.

3.7. Light Gradient Reinforcement Machine Classifier. The Light Gradient Boosting Machine Classifier is a library developed by Microsoft in 2017. LGBM is one of the Gradient Boosting Machine (GBM) types developed to increase the training time performance of XGBoost. XGBoost makes the initial search transversely, while LGBM performs the first search in depth [15]. Increasing the learning rate from default parameters and limiting the number of trees to 150 in improved model studies in LGBM increased the prediction success.

The hyperparameters that are changed when creating an optimized model are shown in Table 3.6. In this study, the LGBMClassifier classification algorithm in the Python programming language lightgbm library was used.

4. Results and Evaluation. There are some criteria to evaluate the classification success of machine learning approaches used on data set. In the calculation of these criteria, the confusion matrix specified in Table 4.1 is used. The confusion matrix indicates the prediction accuracy success in the created model in a 2x2

Table 4.1: Confusion matrix.

Real Class	Class	Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

matrix. It allows the comparison of actual values with the estimates made.

The terms mentioned in Table 4.1 can be explained as follows:

1. True Positive (TP): Specifies the accurately predicted liver disease class value.
2. False Positive (FP): Specifies the inaccurately predicted liver disease class value.
3. False Negative (FN): Specifies the inaccurately predicted non-liver disease class value.
4. True Negative (TN): Specifies the accurately predicted non-liver disease class value.

There are many metrics in the literature that are used to evaluate classification achievements. These metrics and their formulas, which are also used within the scope of this study, are explained in this section.

Accuracy. This metric represents the ratio of accurately predicted liver patients to all values Equ. 4.1.

$$\text{Accuracy }=\frac{TP+TN}{TP+FN+TN+FP} \tag{4.1}$$

Sensitivity. Also known as recall, this metric shows the ratio of accurately predicted liver patients to the sum of correctly predicted liver patients and those incorrectly predicted as not having liver disease. It demonstrates the model's ability to detect true positive cases Equ. 4.2.

$$\text{Sensitivity }=\frac{TP}{TP+FN} \tag{4.2}$$

Specificity. This metric represents the ratio of accurately predicted non-liver disease cases to the sum of all predicted non-liver disease cases. It shows the model's ability to correctly identify true negative cases Equ. 4.3

$$\text{Sensitivity }=\frac{TN}{TN+FP} \tag{4.3}$$

Precision. Ratio of accurately predicted liver patient values to the sum of accurately predicted liver patient values and correctly predicted liver patient values. It shows the ratio of liver patient class values to which liver disease is estimated shown in Equ. 4.4.

$$\text{Precision }=\frac{TP}{TP+FP} \tag{4.4}$$

F-Criterion. Also known as F1-score, this metric is the harmonic mean of Precision and Sensitivity. It provides a single score that balances both precision and recall Equ. 4.5.

$$F1-\text{ Criterion}=2\times\frac{\text{Precision }\times\text{ Recall}}{\text{Precision }+\mathrm{R}ecall}\frac{2\times TP}{2\times TP+FP+FN} \tag{4.5}$$

MAE (Mean Absolute Error). It is the amount of the difference between two continuous variables. Values close to zero perform better. The result values of the machine learning methods applied after the original version and attribute extraction from the data set are given in Table 4.2.

Many datasets can contain many related or unrelated data. The 11 attributes in the dataset were reduced to 4 attributes after the feature selection process. When Table 4.2 is examined, it is seen that the accuracy rates of MLP, SVM, DT, LGBM machine learning algorithms increase after feature extraction. A comparison of the accuracy values after the model refinement is applied after feature selection is given in Table 4.3 and Figure 4.1.

Table 4.2: Comparison of truth values before and after feature extraction.

Classification Algorithm	(%) Accuracy	
	Before Feature Selection	After Feature Selection
SVM	75.21	76.13
MLP	74.32	78.63
DT-CART	75.91	77.04
LR	78.09	77.81
LGBM	76.20	77.95

Table 4.3: Comparison of accuracy values before and after feature selection.

Classification Algorithm	(%) Accuracy		
	Before Feature Selection	After Feature Selection	After Model Optimization
SVM	75.22	76.13	77.87
MLP	74.31	78.63	81.13
DT-CART	75.90	77.04	81.13
LR	78.09	77.80	77.80
LGBM	76.21	77.95	82.12

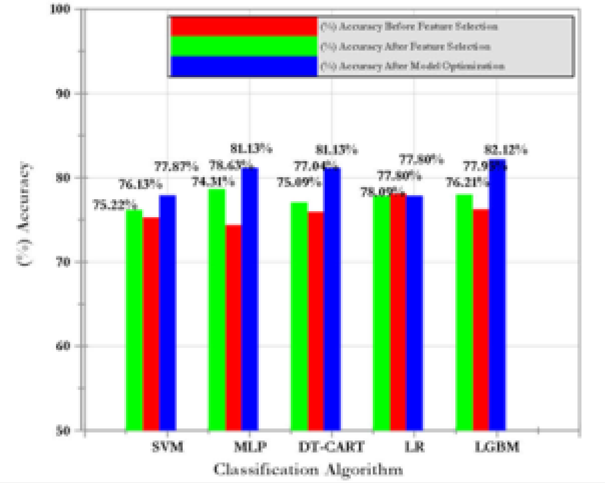


Fig. 4.1: Comparison of accuracy values before, after and model optimization

When Table 4.3 is examined, it is seen that the accuracy rates of all machine learning methods as in the Figure 4.1 increase after feature selection. The performance values of the machine learning methods used after the model improvement application stage after the feature extraction are given in Table 4.4. Figure 4.2 illustrated the performance evaluation of different ML algorithms, in terms of accuracy, sensitivity, specificity, precision and F-criterion.

When Table 4.4 is examined, the maximum diagnostic success rate was got from LGBM machine learning technique with the rate of 82.12%. With a success rate of 81.13%, which is close to the LGBM algorithm,

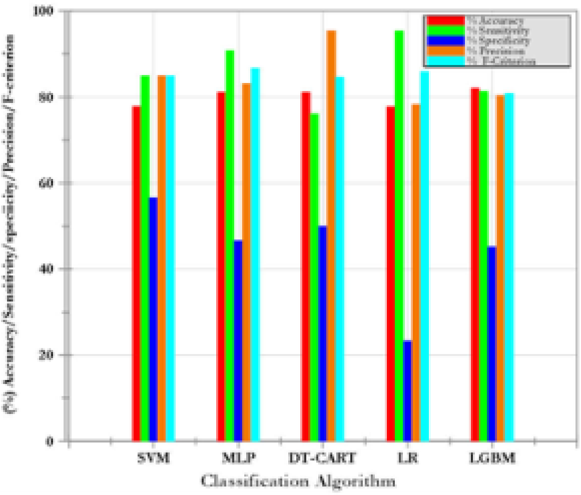


Fig. 4.2: Performance evaluation of different ML algorithms.

Table 4.4: Machine learning methods performance values.

Algorithm	% Accuracy	% Sensitivity	% Specificity	% Precision	% F-Criterion	Mean Absolute Error
MLP	81.13	90.80	46.66	83.15	86.81	0.2
SVM	77.87	85.05	56.66	85.05	85.05	0.22
LR	77.80	95.40	23.33	78.30	86.01	0.23
DT-CART	81.13	76.14	50.00	95.40	84.69	0.25
LGBM	82.12	81.39	45.16	80.45	80.92	0.28

DT-CART and MLP are the second high-performance methods. According to the sensitivity criterion, the highest rate was obtained from the LR method with 95.40%. When the results were evaluated according to the measure of precision, DT-CART was the most successful method with a rate of 95.40%. According to the criterion, MLP is the method that gives the best results with a rate of 86.81%. When the relevant studies in this field were examined, it was seen that LGBM machine learning method was not applied on the ILPD data set before.

Figure 4.3, when the studies evaluated according to MAE were considered, the lowest MAE error rate was obtained. Machine learning is also frequently used in the field of healthcare. With the development of technology, it is aimed to diagnose diseases early by using smart applications as well as traditional diagnostic methods in disease diagnosis. Machine learning algorithms are often utilised in the development of intelligent applications. In this study, the focus is on the early diagnosis of liver failure, which has been seen frequently in recent years and leads to loss of life if not diagnosed in the early stages, with high accuracy. With the feature selection method, it was reduced from 11 attributes to 4 attributes and higher prediction successes were obtained. Models were created on five different machine learning methods used. Model improvements were made to these models and the disease prediction success of the models was increased. In the selection of the model, highest accuracy, precision, sensitivity, and F-Criterion ratios were taken into account. The most successful result was obtained from the LGBM method with an accuracy rate of 82.12%.

5. Conclusion. In this study, we employed Forward Selection (FS) for feature selection and applied various machine learning methods, including Multilayer Perceptron (MLP), Support Vector Machine (SVM), Decision Trees (DT-CART), Light Gradient Boosting Machine (LGBM), and Logistic Regression (LR), to the Indian

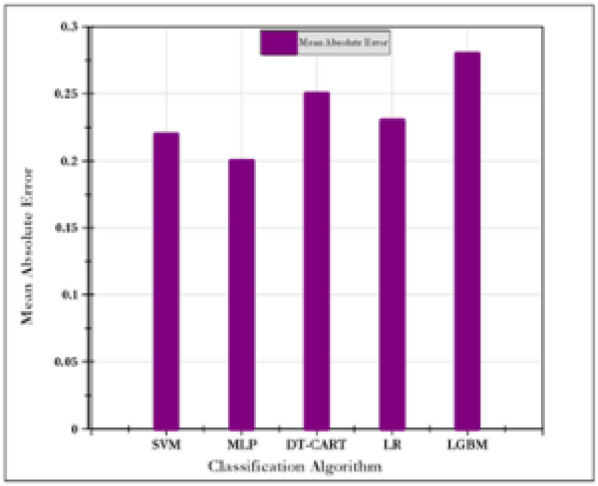


Fig. 4.3: Mean Absolute Error comparison for different Classification Algorithms.

Liver Patient Dataset (ILPD) for the early diagnosis of liver diseases. The diagnostic success of these methods was evaluated before and after feature selection, as well as after model optimization. Among the evaluated methods, the Light Gradient Boosting Machine (LGBM) demonstrated the best overall performance. After feature selection and model optimization, the LGBM achieved the highest accuracy of 82.12%, significantly improving from its initial accuracy of 76.21%. Additionally, the LGBM model showed a balanced performance across various metrics, with a sensitivity of 81.39%, specificity of 45.16%, precision of 80.45%, and F1-score of 80.92%, alongside a mean absolute error of 0.28.

Comparatively, the Multilayer Perceptron (MLP) also showed strong performance, with an accuracy of 81.13% after model optimization, improving from 74.31% before feature selection. However, LGBM's higher accuracy and balanced performance across different metrics make it the preferred choice in this study for early diagnosis of liver diseases.

The results underscore the effectiveness of feature selection and model optimization in enhancing performance of machine learning methods for medicinal diagnostics. Future research could focus on incorporating additional clinical data and exploring more advanced machine learning techniques to further improve diagnostic accuracy and reliability.

REFERENCES

[1] I. Arshad, C. Dutta, T. Choudhury and A. Thakral, , *"Liver disease detection due to excessive alcoholism using data mining techniques",* 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 163-168, 2018.

[2] H. W. Lee, J. J. Sung, and S. H. Ahn, , *"Artificial intelligence in liver disease,"* Journal of Gastroenterology and Hepatology, Vol.36, Iss.3, pp.539-542, 2021.

[3] Mamdouh E. ve Mabrouk M. , *"A Study of Support Vector Machine Algorithm For Liver Disease Diagnosis",* American Journal of Intelligent Systems, 4 (1), 9-14, 2021.

[4] Gulia A. Vohra R. Rani P. , *"Liver Patient Classification Using Intelligent Techniques",* International Journal of Computer Science and Information Technologies), 5 (4): 5110-5115, 2014.

[5] Naeem, S., Ali, A., Qadri, S., Khan Mashwani, W., Tairan, N., Shah, H., ... & Anam, S. , *Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images.* Applied Sciences, 10(9), 3134, 2020.

[6] Priya M.B. Juliet P. L. Tamilselvi P.R." , *Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms",* International Research Journal of Engineering and Technology (IRJET), 206-211, 2018.

[7] Chaurasia, V., Pal, S., , . *Machine learning algorithms using binary classification and multi model ensemble techniques for skin diseases prediction.* International Journal of Biomedical Engineering and Technology, 34(1), 57-74, 2020.

[8] Gertz, M.; Adams, D.; Ando, Y.; Beirão, J.M.; Bokhari, S.; Coelho, T.; Comenzo, R.L.; Damy, T.; Dorbala, S.; Drachman, B.M.; et al. , *Avoiding misdiagnosis: Expert consensus recommendations for the suspicion and diagnosis of transthyretin amyloidosis for the general practitioner.* BMC Fam. Pract., 21, 198, 2020.

[9] Ahmed, S. T., Kumar, V. V., & Kim, J. , *AITel: eHealth augmented-intelligence-based telemedicine resource recommendation framework for IoT devices in smart cities.* IEEE Internet of Things Journal, 10(21), 18461-184685, 2023.

[10] Eltanashi, S., & Atasoy, F. , *Proposed speaker recognition model using optimized feed forward neural network and hybrid time-mel speech feature.* ICATCES 2020 Proceeding Book, 130-140.

[11] https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset).

[12] Koike, H.; Katsuno, M. , *Transthyretin Amyloidosis: Update on the Clinical Spectrum, Pathogenesis, and Disease-Modifying Therapies.* Neurol. Ther. 9, 317–333, 2020.

[13] Tozza, S.; Severi, D.; Spina, E.; Di Paolantonio, A.; Iovino, A.; Guglielmino, V.; Aruta, F.; Nolano, M.; Sabatelli, M.; Santoro, L.; et al. , *A compound score to screen patients with hereditary transthyretin amyloidosis.* J. Neurol. 269, 4281–4287, 2022.

[14] Manjunatha, S., Swetha, M. D., Rashmi, S., & Subramanian, A. K. , *Convolutional Neural Network-based image tamper detection with Error Level Analysis.*, 2024.

[15] Di Stefano, V.; Thomas, E.; Alonge, P.; Giustino, V.; Pillitteri, G.; Leale, I.; Torrente, A.; Pignolo, A.; Norata, D.; Iacono, S.; et al. , *Patisiran Enhances Muscle Mass after Nine Months of Treatment in ATTRv Amyloidosis: A Study with Bioelectrical Impedance Analysis and Handgrip Strength.* Biomedicines, 11, 62, 2022.

[16] Subashchandrabose, U., John, R., Anbazhagu, U. V., Venkatesan, V. K., & Thyluru Ramakrishna, M. , *Ensemble Federated learning approach for diagnostics of multi-order lung cancer. Diagnostics* 13(19), 3053, 2023.

[17] Luhadia, G., Joshi, A. D., Vijayarajan, V., & Vinoth Kumar, V. (2023, April)., *Fusion of Variational Autoencoder-Generative Adversarial Networks and Siamese Neural Networks for Face Matching. In International Conference on Frontiers of Intelligent Computing: Theory and Applications, Singapore: Springer Nature Singapore.*,231-241, 2023.

[18] Dhiman, G., Vinoth Kumar, V., Kaur, A., & Sharma, A. , *Don: deep learning and optimization-based framework for detection of novel coronavirus disease using x-ray images. Interdisciplinary Sciences: Computational Life Sciences*, 13, 260-272, 2021.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jul 7, 2024
Accepted: Sep 18, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1116–1131, DOI 10.12694/scpe.v26i3.4262

CLOUD COMPUTING-BASED DYNAMIC RESOURCE ALLOCATION, CC-DRAM, FOR ONLINE LEARNING PLATFORM

YU HU[*], HUI WANG[†], JIANGTING TANG[‡], AND JIE YANG[§]

Abstract. A cloud classroom is a new type of online education that has recently evolved within the framework of the Internet and education. Learning in a cloud classroom means students access course materials and goals online, collaborate with instructors and peers, and construct their knowledge base via the Internet. There is an insufficient individualized suggestion module and no way to alleviate information overload, which are features of the conventional cloud classroom model of instruction. Hence, this paper proposed a Cloud Computing-based Dynamic Resource Allocation Model (CC-DRAM) to improve content delivery and increase resource allocation in online learning. Consequently, the CC-DRAM operating under the customized recommendation system is used in this research. The system uses a collaborative filtering recommendation algorithm to enhance cloud work scheduling, learn users' preferences, and provide better suggestions. It also allows for the integration and integrated management of different resources through technologies like distributed storage, virtualization, and networking. Based on experimental analysis of the CC-DRAM platform, which provides 24/7 access to digital materials for students and educators, we can now create individualized lesson plans that students and instructors may read, download, print, and share. In this proposed method, the scalability of distributed storage, user satisfaction, performance, the effectiveness of collaborative, and resource allocation metrics are analyzed and compared to the existing method; the values are gradually increased by the ratio of 97.8%, 98.2%, 99.34%, 96.12%, 98.41% respectively.

Key words: Cloud Computing, Resource Allocation, Online Learning

1. Introduction.

Cloud computing has revolutionized several sectors, including education, by offering scalable, efficient, and adaptive resources [1]. A cloud classroom is a concept that has emerged in this dynamic setting. Teachers and students in an online classroom collaborate and share educational materials over the web [2]. According to this streamlined access to materials, peer engagement, and instructor support, students are able to build their own knowledge bases. Additionally, it enables students to learn in an interactive and cooperative setting [3]. In evaluation, conventional cloud school rooms have numerous problems, particularly with offering customized gaining knowledge of experiences and dealing with facts overload. Due to its lack of sophisticated methods for offering personalized records, the conventional method often leaves students feeling filled with stuff this is both superfluous or unneeded [4]. Some have proposed the use of the CC-DRAM to address these troubles. By streamlining the allocation of assets and the dissemination of information presented by means of on-line learning systems, this advanced technique pursuits to decorate the educational enjoy as a whole [5].

Advanced cloud computing technology such networking, virtualization, and disbursed garage permit CC-DRAM to efficiently combine and control a various range of educational content material [6]. The core of CC-DRAM is a collaborative filtering recommendation algorithm that powers a custom designed advice device [7]. The system can research user alternatives and offer personalised hints, so the studying enjoy can be customized to suit each person's desires and tastes. Improving cloud work scheduling requires a collaborative filtering advice mechanism [8]. An critical a part of those enhancements is this approach's guarantee of green

[*] School of Economics and Management of Hunan University of Science and Engineering, YongZhou, HuNan,425199, China (yzlmg123@163.com)

[†] School of Economics and Management of Hunan University of Science and Engineering, YongZhou, HuNan,425199, China (daincywang@126.com)

[‡] School of Economics and Management of Hunan University of Science and Engineering, YongZhou, HuNan,425199, China (tjt19870929@163.com)

[§] School of Information Engineering of Hunan University of Science and Engineering, YongZhou, HuNan, 425199, China, (Mryj1975@163.com)

useful resource distribution in keeping with consumer behavior and preferences [9]. In addition to mitigating the difficulty of facts overload, dynamic resource allocation improves the efficacy of content distribution by way of ensuring that students gain enticing and relevant contents [10].

Experiment results exhibit that the CC-DRAM platform efficaciously offers ongoing get entry to to digital pedagogical materials [11]. The ability to view, download, print, and share these resources at any time helps create a more adaptable and responsive learning environment for both students and instructors [12]. Individualized learning paths that address each student's unique needs are made possible with the help of CC-DRAM's facilitation of course plan preparation [13]. Cloud computing (CC) has grown rapidly since students realized its advantages over conventional IT systems. This paradigm enables distributed computing systems, data management, and computing resources via scalable networks, data processing centers, and web services [14]. Thus, this technology is driving the distributed computing revolution and accelerating commercial and public platform development. Users must negotiate and sign a service level agreement (SLA) to access items in utility computing. After signing a contract for computing commodities, users and the CC system (via service maintenance) must cooperate [15]. This marketing approach legally demands CC systems to maintain QoS, hence the internal architecture must dynamically monitor and adjust to demand. The range of innovative underlying technologies, such as web services and micro services, virtualization, dynamic resource allocation, and service farms, has enabled dynamic service delivery regardless of user demand.

Contribution of this paper.

1. Designing the Cloud Computing-based Dynamic Resource Allocation Model (CC-DRAM) to improve content delivery and increase resource allocation in online learning.
2. Introducing the collaborative filtering recommendation system to recognize and recommend each student's choices may meet their needs. This personalized approach aims to improve online classes' inability to customize learning.
3. The paper aims to reduce information overload in online classes, and the proposed CC-DRAM offer educational materials without overloading pupils. This method makes learning more flexible and responsive and improves information delivery.
4. CC-DRAM employs distributed storage, virtualization, and networking to manage instructional resources efficiently. This facilitates the early creation of individualized lesson plans and ensures constant access to digital resources. This method improves online education's overall efficacy by decreasing overwhelming material.

2. Related works. Allocating assets dynamically is critical in cloud computing for optimizing pace, however it faces challenges with power consumption, fault tolerance, and service quality. The cloud computing enterprise faces each and every one of these difficulties directly. In mild of these issues, this paper gives fashions and strategies to decorate cloud computing overall performance. A famous method that complements fault tolerance and strength intake at the same time as lowering makespan and digital system costs is the Spacing Multi-Objective Antlion technique, or S-MOAL. The Adaptive Multi-Objective Teaching-Learning Based Optimization (AMO-TLBO) algorithm can stability masses and store charges by means of adjusting to purchasers' ever-changing expectations. These methods improve virtual machine balance and offer efficient useful resource allocation through tackling scheduling troubles via multi-objective optimization.

The cloud's dynamic resource allocation is considered one of its quality capabilities. However it has considerable problems with power use, fault tolerance, and carrier best. Finding a solution with a purpose to decorate cloud overall performance even as also fixing these vital worries became vital. To higher and extra speedy respond to customer call for for sources, it introduces a model for dynamic resource allocation. To similarly lessen the makespan and rate associated with digital machines, it suggests a multi-objective seek algorithm called the Spacing Multi-Objective Algorithm (S-MOAL) [16]. Its results on electricity utilization and fault tolerance had been additionally investigated. According to the outcomes of the simulation, our method outperformed the PBACO, DCLCA, DSOS, and MOGA algorithms, mainly whilst thinking about makespan.

The cloud data center allocates resources for multiple fine computational granularity jobs, a non-polynomial complete issue. The needs of customers and the capabilities of apps are subject to constant change. It provide a dynamic resource allocation technique in Cloud computing using an Adaptive Multi-Objective Teaching-Learning Based Optimization (AMO-TLBO) algorithm to close the gap between the ever-changing customer requirements and the available service infrastructure [17]. Adaptive teaching factors, tutorial training, self-motivated learning, and the number of instructors are all introduced by AMO-TLBO to enhance the capabilities for exploration and exploitation. Minimizing makespan and expense while optimizing utilization by load balancing among virtual machines are the aims of AMO-TLBO. Machine learning and AI have provided practical solutions to complicated issues such as energy efficiency, workflow scheduling, video gaming, and cloud computing. Combining machine learning and cloud computing techniques improves cloud data center performance compared to existing examiner solutions. It also aids virtual machine migration depending on network congestion and bandwidth availability. It uses machine learning categorization to show improvements in dynamic load allocation, work scheduling, energy optimization, live migration, mobile cloud computing and cloud security [18]. Machine learning algorithms are popular analytical methods that help computers find patterns and simplify learning. Introduction, motivation, preliminary analysis, framework for cloud-machine learning integration, best practices for incorporating machine learning in cloud computing, and work target make up the paper. The analysis also discusses machine learning-based cloud services and AI in cloud computing platforms. This complete analysis of machine learning techniques and cloud computing gives researchers with insightful and essential resources.

When compared to cloud computing, edge computing is capable of efficiently resolving the issue of excessive latency that is present in cloud gaming. However, there are still a number of obstacles that need to be overcome to maximize the performance of the system. Unpredictable gaming demands might overburden servers and networks however, player movement makes the system dynamic. A tradeoff between fairness and latency has been generally disregarded, despite the fact that prior work has investigated game fairness and latency independently to enhance the Quality of Experience (QoE). Optimization also faces network and computational load balancing constraints [19]. It present an adaptive resource allocation technique for a dynamic gaming system that makes use of Deep Reinforcement Learning (DRL). This strategy takes into consideration latency, fairness, and load balancing all at the same time. This method solves difficult multimodal reward issues better than standard optimization and classical reinforcement learning algorithms, according to experiments.

Cloud computing is a milestone in commercial distributed computing and offers promising possibilities. It uses virtualization to aggregate large resources from disparate locations for unified administration and consumption. Optimization of virtual machines allocation improves resource usage, reduces expenses, and saves computation time. A multiobjective optimization strategy for dynamic resource allocation for multivirtual machine distribution stability is proposed in this paper. Each application load's present and future anticipated statistics are used to calculate virtual machine relocation costs and stability [20].

The operational model is laid out to illustrate the interdependencies between the variables. This study takes an explanatory approach by surveying 143 college students utilizing cloud-based e-learning. Data quality is ensured using descriptive statistics and validity tests, while hypothesis testing utilizes Mediated Regression Analysis. Social influence has an unclear direct effect on cloud-based e-learning, although relative advantage and user happiness have a beneficial effect. Adopting cloud-based e-learning is complex, as behavioural intention does not mediate the interactions as imagined. This research helps fill gaps in understanding how educational institutions evaluate new technologies by shedding light on the complex dynamics in the decision to use cloud-based e-learning [22].

The Cloud computing-based online and offline hybrid teaching resource-sharing method is constructed with a three-tiered cloud platform for sharing virtual and physical educational materials; this platform will employ the fuzzy neural network model and the Tucker decomposition technique to combine the features derived from the resources. The next step is to develop a paradigm for sharing teaching resources that utilizes layered agent technology. This model will allow for the combination of online and offline teaching resources. This paper presents a design method that consistently achieves a resource request success rate of over 80%, a maximum data sharing of over 98%, and a sharing time of less than 1 s. Experimental results confirm that this method has a high sharing efficiency [23].

Mobile learning technologies (MLTs) allow for discovering correlations and patterns relevant to adaptation via examining extensive datasets that include markers of student behaviour, performance, and engagement within online platforms. By looking back at factors, including teachers' level of technical competence, their level of motivation, and their capacity for self-regulation, the OLAMLT framework may provide tailored suggestions.

The project aims to bridge the gap between the need for flexible learners and the lack of resources to cultivate this important quality by promoting focused educational interventions. The author hopes that by strengthening online learning systems, it can better withstand and recover from future shocks, such as pandemics or other unexpected obstacles. Thanks to this study, education technology and pedagogy have taken a giant leap forward, which adds to the continuing push for a more robust and flexible online learning environment [24].

An online learning management system that can be accessed anytime and anywhere makes managing learning materials, course administration, and digital interaction with students easier. These are some ways a cloud-based learning management system can improve access and quality of education. This is why this research was carried out. Quantitative approaches were used in this study. This approach is a means of gathering testable facts and numerical information. Student questionnaires were distributed to gather data. In addition, you will have access to the data in Excel format, which may be analyzed using SPSS according to the findings of the questionnaire distribution. According to the study's findings, a cloud-based learning management system can potentially increase the quantity and quality of educational opportunities. Beyond that, a cloud-based learning management system may enhance instructors' competitive performance in the classroom [25].

This research used various statistical methods to delve deeply into the complex issues surrounding cloud computing and its consequences. Several statistical methods were used to examine the data in this extensive study. These included t-tests, descriptive statistics, and analysis of variance (ANOVA). This study highlights the significance of skillfully addressing localization, script support, and linguistic differences in disseminating Arabic information. This research highlights the significant possibilities of cloud computing to improve online education's effectiveness and user experience [26].

Agricultural professionals may use a high-efficiency online learning platform in the cloud that employs real-time streaming analysis to track the network string flow as users watch videos and dynamic allocation to get the most out of each server. Users may stay inspired while enrolled in top-notch virtual classes. It wants to build cloud-based course materials tailored to agricultural knowledge to enhance user motivation and learning effectiveness. This study used satisfaction surveys and the UTAUT model to evaluate the research results further. The model determines whether the positive effect of users' performance expectation, effort expectancy, social influence, and enabling factors on perceived satisfaction and usefulness is significant. Based on our verification of this fact, this study deduced that digital learning may greatly benefit the agricultural community [27].

Many primary critical success factors (CSFs) and subfactors affect sustaining success in M-learning. This research evaluates and ranks several primary and secondary components of CBML. In both crisp and fuzzy settings, the primary components and subcomponents of CBML were examined and modelled using approaches based on fuzzy analytic hierarchy processes (FAHP) and analytical hierarchy process-group decision-making (AHP-GDM). Higher education institutions must address these primary and secondary aspects to achieve their goals in the teaching-learning system and implement sustainable M-learning [28].

Based on the survey, there are several issues with existing models in attaining high scalability of distributed storage, user satisfaction, performance, the effectiveness of collaboration, and resource allocation. Hence, this study proposes a Multi-objective Optimization Genetic Algorithm(MOGA) to overcome existing problems. The simulation results suggest that the MOGA Virtual Machine Distribution approach has a longer stability time than the genetic algorithm for energy saving and multi-virtual machine redistribution overhead. To address this problem, the MOGA multi-objective optimization dynamic resource allocation approach is introduced for virtual machine distribution. The findings of the proposed method simulation show that these strategies are more effective in managing cloud resources reliably and efficiently than the current methods.

3. Proposed Method.

The dynamic and effective gaining knowledge of environments has been ushered in by the combination of cloud computing within the constantly converting subject of training technology. In this studies, a entire framework for enhancing the distribution of instructional substances and aid optimization the usage of smart cloud-based structures is introduced. With the use of cloud computing, information analytics, and sensible allocation algorithms, the cautioned technique seeks to absolutely transform the way that training is taught and learnt. Cloud computing's feature is to assist instructional establishments store charges and deal with their essential enterprise. While vendors pay for the prices of supplying hardware and software, instructional establishments are best paid for the services and sources they use, consisting of computer and

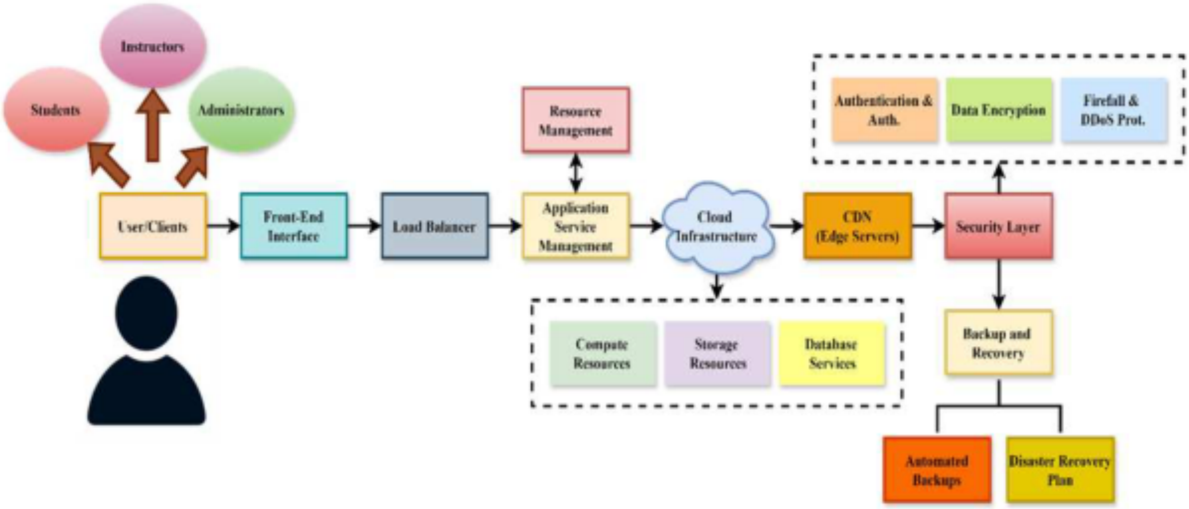


Fig. 3.1: Cloud Infrastructure for Online Learning Platform

storage sources, specialised clinical software program structures, and lecture substances.

A complicated digital atmosphere with many components is interacted with by users/customers. A front-stop interface, typically an internet or cellular app, permits users consisting of students, instructors, and directors to get entry to the machine. Application servers oversee content material transport and scalable instances; a load balancer makes certain that traffic is shipped correctly amongst them. Within a cloud structure, those servers run applications that employ storage, computation, and database services. A Content Delivery Network (CDN) with strategically positioned area servers optimizes content material delivery for advanced overall performance. Authentication, encryption, and defence towards threats like dispensed denial of provider assaults are all a part of a safety layer's activity to maintain person facts and their privateness secure. Data loss and gadget outages may be prevented by backup and restoration procedures, consisting of computerized backups and catastrophe recovery plans. All matters taken into consideration, this complicated layout locations an emphasis on dependability, scalability, protection, and person experience, permitting customers to have interaction with the platform without any hitches whilst yet ensuring the safety and accessibility of their information shown in Fig.3.1.

$$U_j j = J(-N < T_p < -1)U_p j, p + J(T_p < -N)u^J u, n+1 \tag{3.1}$$

Users' satisfaction or utility levels throughout defined time periods T_p are denoted by the equations $U_j j$ and U_p, respectively in Equ.3.1. The notation N denote distinct time intervals that allow for modifications to be made to the allocation of resources. The concept of iterative improvement and updating of consumer needs and satisfaction (u) across subsequent time periods (n+1) is implied by the phrase u^J u,n+1.

$$D_o qw = GP(-Qa < Ws_p < -1)CU_p pk, se + O(WdT_p < -W)fu^J mu, nw + 1 \tag{3.2}$$

In the Equ.3.2, D_o and qw, the demand and quality of service weights are represented. For a certain process CU_p, the range that is appropriate of workload Ws_p is represented which guarantees that it remains within optimum bounds. The term GP $(-Qa < Ws_p < -1)$ is used to describe the efficiency of resource utilization or computational utility for process p, which is affected by parameters pk and se. To make sure the distribution of workloads over time Ws_p stays below a threshold W, the optimization function is shown by the component $WdT_p < -W$.

$$QV_q(u) = \frac{1}{\forall_e d(u)} \sum_{n=1}^{P} Q + e \sum_{f}^{e} r + s * (w + at) - (r_s(u)) \tag{3.3}$$

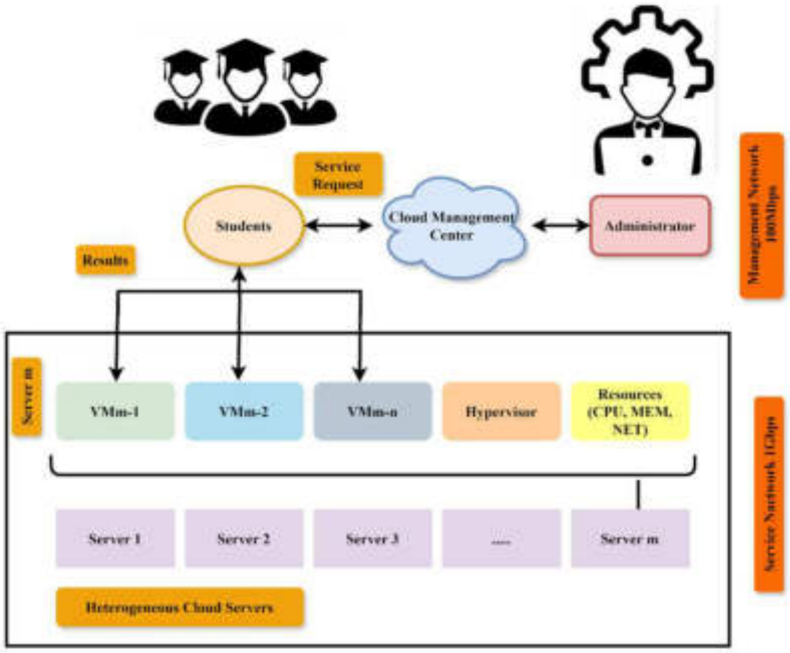


Fig. 3.2: Cloud Computing Approach for Aggregated Data and Information Mining Across Many Platforms

For user u, the quality value is represented by the Equ.3.3, $QV_q(u)$. Every user's overall demand or resource needs are represented by the unit of measurement $\forall_e d(u)$. These measurements are aggregated across P periods or jobs by the summation n=1. In the layered summation inputs from specific resources or tasks are accounted for, and Q represents the basic quality rating inside the summation. e is a scaling factor. With constants r and s, workload w and allocation time at and residual or unsatisfied demand for user, the effective allocation of resources may be represented by $(w + at) - (r_s(u))$.

Cloud computing's statistical Big Data and information mining paradigm, including the Cloud Management Centre and VM servers, is graphically shown in Fig.3.2. Connecting several virtual machines (VMs), the CMC oversees user access for those wishing to utilize the remote statistical analysis services. The CMC also links the various VM servers and handles the VM states. Virtual machine servers are executing user programs. For statistical analysis, please be informed that our cloud computing system provides one virtual machine per user. When processing data offline, administrators have the option of pre-assigning user virtual machines to certain servers; when users are getting statistical analysis findings, these VMs may be maintained in standby mode rather than terminated. Additionally, a management network and a user network are separated to control the networks and provide consistent services. The management network is where user access and VM state management take place. As a result of the ease and effectiveness it provides in managing centralized desktop administration, businesses are increasingly turning to cloud management centres. These centres employ a virtualization approach based on virtual desktop architecture to provide cloud services. This is why we build the cloud-based statistical services into a heterogeneous platform for massive users, capitalizing on the SaaS and DaaS concepts.

$$Q(E) = f^{\frac{-\alpha(g)}{T(U)}} + T(u+1) * \partial T(u) * \partial(F(u)) \tag{3.4}$$

The system's efficiency or quality is represented by the Equ.3.4, Q(E). The influence of a scaling factor $\frac{\alpha(g)}{T(U)}$ across the whole resource time T(u+1) is represented by the expression suggesting an exponential connection in which efficiency declines as load increases. The iterative refinement process is described by the expression

T(u+1), where $\partial(F(u))$ is the time in the following iteration, F(u) is the change in resource time, and $\partial T(u)$ is the change in user feedback or performance metrics.

$$\partial^q(r+1,s,u) = \alpha^q(r,s,e) + T(v) * (1+pqt) \tag{3.5}$$

The modified allocation of resources for the following iteration ∂^q is represented by the equation r+1,s,u, which takes into account the type of resource and the use. The present allocation effectiveness, affected by parameters r, s, and the environment e, is denoted by the expression. The time needed for the process is adjusted by a factor in the formula T(v), where 1+pqt are parameters is an adjustment factor.

$$G(T_B) = \sum_{j=1}^{p}(FG)T_2b^J * (1 + C_F G(H+1)) + (R_f g(k-1)) \tag{3.6}$$

Contribution from every user or jobs up to p are aggregated in the Equ.3.6. The basic gain (FG) T_2b^J a user-specific component is represented by the phrase $1 + C_F G(H+1)$. This gain is adjusted using the equation $R_{fg}(k-1)$ using the factor of correction modified. The next step is to include a residual gain component from the previous cycle's $G(T_B)$.

An interactive getting to know tool, the Cloud Classroom lets in for clean communication among instructors and their students. Digital sources such as readings, downloads, prints, and sharing opportunities are made to be had to students thru an individualized recommendation machine that considers their specific learning necessities. At the identical time, teachers make use of technology for coping with resources to preserve tune of scholar paintings and verify their development. Cloud Classroom Dynamic Resource Allocation Module, or CC-DRAM, is the brains at the back of the Cloud Classroom; it ensures scalability and ultimate resource allocation. Collaborative filtering methods are used in this module to enhance mastering consequences and interactivity. The platform includes assessment measures that measure scalability, consumer happiness, performance, and collaborative effectiveness, in addition to disbursed storage for green records control and retrieval. To take things a step further in phrases of customization, purchaser possibilities are also considered. By offering a flexible, collaborative, and expandable space that promotes efficient instruction, the Cloud Classroom ultimately benefits educators and their students is shown in Fig.3.3.

$$U_j = \sum_{k=j}^{q_r} \forall^{q_1} * \frac{Y^(q1-1) * f^{-\bigcup S}}{\Delta q_j} + PLR\frac{R}{s+q} \tag{3.7}$$

The U_j combines contributions from state k=j in this Equ.3.7. The $Y^(q1-1)$ $f^{-\bigcup S}\Delta q_j$ represents a performance metric where PLR are operates modulated by q_r and the complementary value of s, divided by the gradient $\forall^{q_1}$. The term stands for a universal quantifier function raised to $\frac{R}{s+q}$.

$$PLR([\frac{R}{(s+q)}]) = Q[min(u_j) \ni \frac{O_b}{O}] + (1+sfg) \tag{3.8}$$

The Equ.3.8 in which the variables R stand for resources and the parameters s+q are those that influence the distribution $\frac{R}{(s+q)}$ of those resources. Q is a factor of quality that takes into account the minimal utility for user u_j, conditional on the ratio of observed gain b to optimum benefit $\frac{O_b}{O}$, and the 1+sfg.

$$DS(pe) = \sum_{(j,k)\ni(p,\bigcup,j)}^{r} D_{j,k}(u+1) + (vh+kpr) \tag{3.9}$$

Up to the limit r, the addition of all contribution from pairs $D_{j,k}$ that are components of the combined set $\sum_{(j,k)\ni(p,\bigcup,j)}^{r} D_{(j,k)}$ is represented by the Equ.3.9, DS(pe). The dynamic distribution matrix D for the pairings u+1 updated for the following iteration u+1 is indicated by the phrase $D_{j,k}$.

The User Interface Layer facilitates communication between the educational platform's users (students) and its administrators (teachers) by way of a user-friendly interface. It has features that are customized to

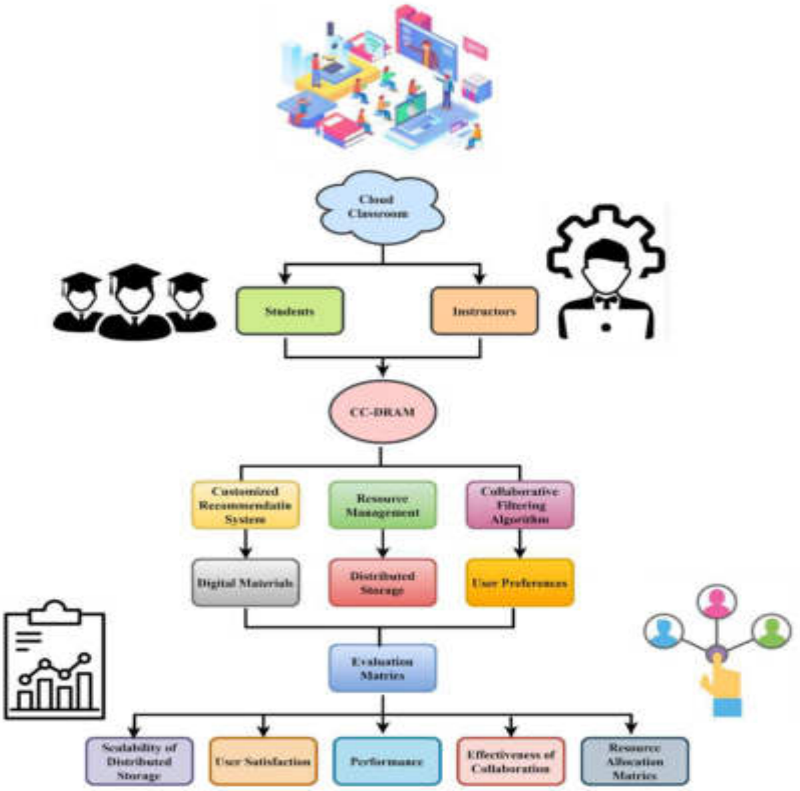


Fig. 3.3: Cloud Computing-based Dynamic Resource Allocation Model

cater to the different requirements of both parties. Under everything is the Application Layer, which is where all the tools for managing content and learning are located. To make sure that instructional materials are easily accessible and organized, here is where the features of content distribution and course administration are coordinated. Incorporating features such as CDN, load balancing, and resource management, the Middleware Layer connects various levels. For steady overall performance and scalability, this sediment optimizes the allocation and use of sources. The platform's functionality and statistics management are supported by way of the VMs, boxes, and database offerings which can be hosted in the Service Layer. Computing, garage, and networking resources make up the Cloud Infrastructure Layer, that's the backbone of the machine and the basis for how the platform functions. Scalability and dependability are guaranteed by CC-DRAM's use of cutting-edge cloud computing technologies, which dynamically manage and allocate computing resources like CPU, memory, and storage according to real-time demand. While improving the user experience and lowering operating expenses, this dynamic allocation ensures constant performance even during high-demand periods. The strong data management approach lies at the heart of CC-DRAM's efficacy. The approach incorporates effective methods for gathering, storing, and analyzing massive amounts of educational data. Data is protected, readily available, and effectively managed using CC-DRAM's scalable storage solutions like data lakes and databases. With this all-encompassing data management strategy, online learning platforms may meet their varied demands, such as keeping track of students' information and course materials and assessing their activities and the results of their lessons. The CC-DRAM system allows online learning platforms to easily adjust to different workloads by combining dynamic resource allocation with smart data management. This connection guarantees the effective and reliable delivery of instructional information while improving resource usage. A vital resource for delivering first-rate instruction, the model can adapt to changing demand and efficiently manage massive amounts of information. Online learning platforms may improve the quality of education

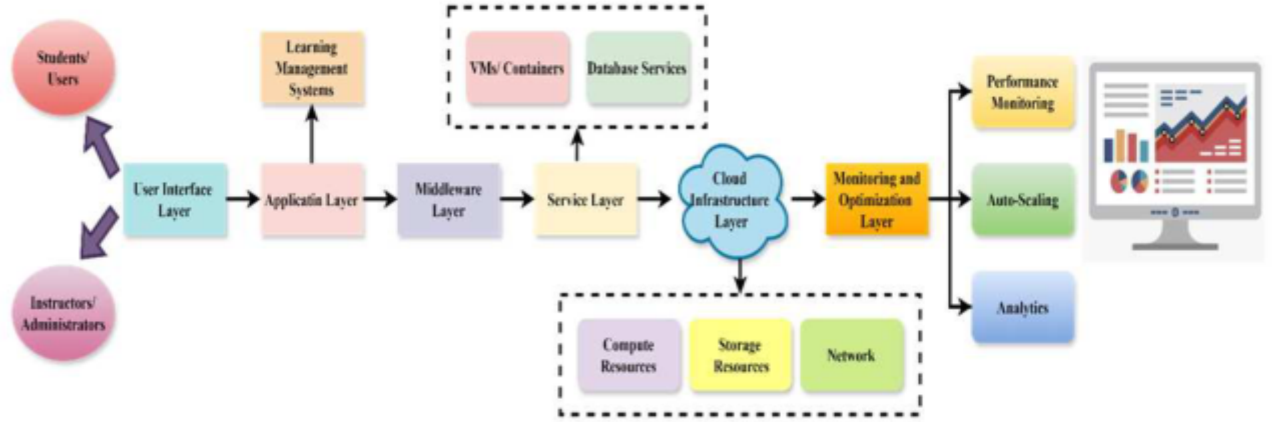


Fig. 3.4: Block Diagram of Cloud Computing-based Dynamic Resource Allocation Model

they provide students by being more scalable, reliable, and cost-efficient using CC-DRAM. Analytics, auto-scaling, and overall performance tracking make up the Monitoring and Optimization Layer, which keeps the device strolling smoothly and effectively at the same time as additionally enhancing the platform's overall performance and consumer experience is shown in Fig.3.4.

$$G_j^e(u) = \sum_{h\forall lhest, k\neq 1}^{p} spef_k * H(u) + \frac{N_k(u) * R_f(N)}{k_j k + e} + (uk + 2) \tag{3.10}$$

The terms up to p in the Equ.3.10 represent the total contributions from $h\forall lhest, k \neq 1$ omitting the values. A particular effectiveness factor $spef_k$ increased by a function of utility H(u) is represented by the expression $G_j^e(u)$. The normalized sum function of the resource is represented as the fraction $\frac{N_k(u)*R_f(N)}{k_{jk}+e)}$, where composite parameter is a setting factor.

$$U_s(u+1) = sf_g(k+1) * H_j^e(k+1) + f * e_k(g+h) \tag{3.11}$$

At iteration Analysis of the scalability of distributed storage, the efficiency of storage allocation for user $U_s(u+1)$ is denoted by Equ.3.11. The interplay of components f and H_j^e regarding g and h parameters, affecting the usefulness of storage, is denoted by the expression $f * e_k(g+h)$.

$$Z_e^q d + e) = (_w^E)sd + (e^f)(u + 2k) + (_g^f e)c(q + w) \tag{3.12}$$

Analysis of user satisfaction Z_e^q and the system's efficiency d+e affected by factors e^f are taken into consideration by Equ.3.12. The last term, $(_g^f e)c(q+w)$, represents the consequences of collaboration between users and resources, which are affected by the parameters sd.

A resource allocation assistance for education is the essential component of the architecture. Information is sent to the resource allocation system via this component, which then releases and allocates resources based on the demand from the school. components of the course materials, class schedules, and more. More "confidence" in the process of changing resource allocation may be assumed, resulting in finer granularity. For instance, aware that there is a class that requires a lot of resources. We pre-allocate resources based on peak-hour demand plus a certain buffer to ensure service quality. So, the resources sit mostly unused outside of the peak period, which is dependent on the school's schedule, but the cloud provider will still take payment. With the system being able to pre-allocate the expected resources before the event, schools would not have to worry about wasting money on idle resources and release resources throughout the schedule's blank spots. Fig.3.5 shows that CC-DRAM approach works just well for both in-person and online class times. When thinking about students using resources and services outside of certain times, the approach would have to change.

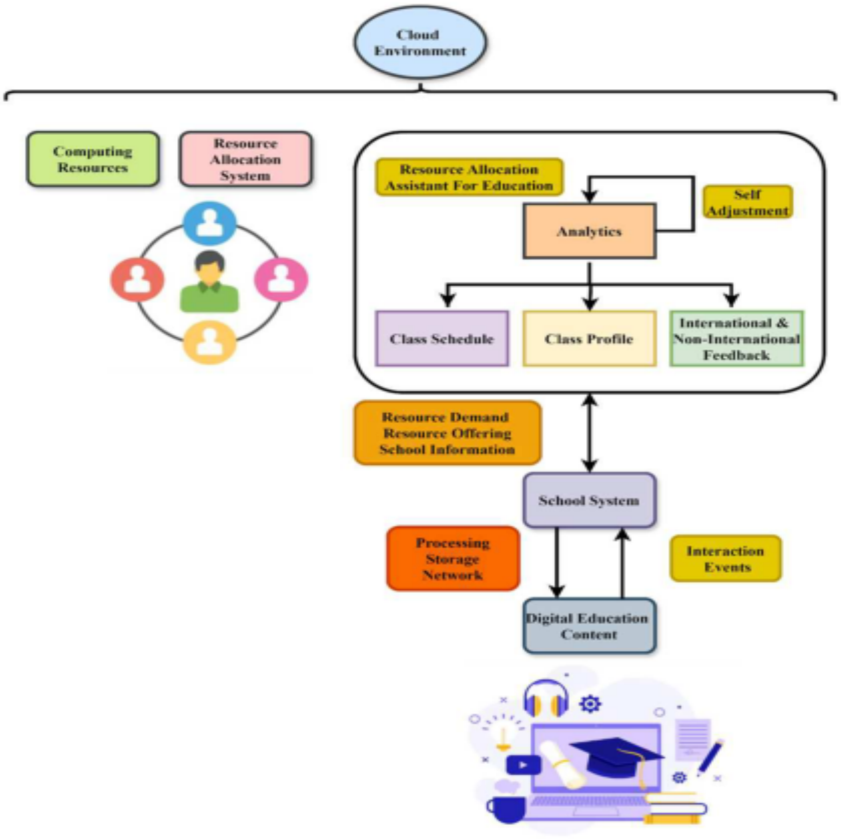


Fig. 3.5: Framework for the Intelligent Distribution of Educational Resources

$$V(u) = V(E_{initiate}, V) + (r + stu_{end}) + e^{-iwt}) \tag{3.13}$$

At the time of initialization in Equ.3.13, the initial level analysis of performance is represented by V(u). The total effect of allocating resources $V(E_initiate, V)$, using the system $(r+stu_{end}$, and a variable representing system stability is accounted for by the expression e^{-iwt}.

$$Z_0.1 + Z_0.2 = U_0 + Xl_e + (U_1 + JR_s) \tag{3.14}$$

Analysis of effectiveness of collaborative separate sources are represented by the Equ.3.14. The basic level of cooperation, denoted as U_0, is affected by the influence Xl_e that is brought about by forces outside of the system $(U_1 + JR_s)$. Furthermore, the adjusted degree of cooperation, denoted by $Z_0.1 + Z_0.2$ component that reflects the combined efforts of collaborators, is also included.

$$A_{01} + B_{03} = E_f(u + 1) + (rsf + (q * w)) \tag{3.15}$$

Analysis of resource allocation within the CC-DRAM system is characterized by Equ.3.15. In this context, the allocation of assets from various sources is represented by $A_{01}+B_{03}$, which adds to the total pool of resources. At the next iteration u+1, the development of the resource environment is represented by $rsf + (q * w)$.

In summary, when applied to educational contexts, the suggested strategy offers a novel way to distribute and allocate resources. Utilizing the capabilities of cloud computing, managers may dynamically distribute resources according to current demand, guaranteeing efficient and effective use. To maximize performance and

Algorithm 8 Code Snippet for an Educational Data-Driven Dynamic Resource Allocation Scheme

```
Input: Class Timetable, Class Outline, Device Connections
Output: Updated Class Outline
1 class – choose Class (Class Timetable)
2 load – get Desired Load (Class Outline)
3 service Delay- get DesiredServiceDelay (Class Outline)
4 service Resources (class.start Time, ServiceDelay, ClassOutline)
While request for resources do
Observe user interaction and resource utilization
If request changed then
Alter resources
Update (Class Outline)
Free resources

Return classOutline
```

decrease waste, the framework uses advanced analytics to forecast useful resource wishes with pinpoint accuracy. Additionally, students have clean get right of entry to to educational materials through smart distribution, which encourages active participation and teamwork in the school room. Using technology to enhance instructional consequences and alter to the converting requirements of each college students and teachers is an ongoing undertaking, however this new technique takes a massive soar ahead.CC-DRAM's operation is guided by a personalized recommendation system that customizes the distribution of resources according to the preferences and actions of the user. The program determines the best resource allocation using machine learning techniques to forecast user demands and adapt to unique courses. With its adaptive performance and personalized content distribution, this smart system improves the learning experience despite fluctuating workloads. The combination of CC-DRAM's dynamic resource allocation, powerful data management, and personalized recommendation system makes it a critical tool for creating a better educational experience. Online learning systems may become more reliable, scalable, and cost-efficient using CC-DRAM, which means better education for students.

4. Result and Discussion. Several aspects of the CC-DRAM will be analyzed in this paper to ascertain its potential usefulness in enhancing online education. To determine whether CC-DRAM's distributed storage technology can handle the increasing demands of users and the massive volumes of instructional materials, analysts are examining its scalability. The level of satisfaction of users is measured by the feedback they give on personalized suggestions and their overall experience with the site. Analysis of performance primarily aims to determine the system's responsiveness, resource utilization, and availability. The amount of user engagement and the material's relevance are factors in determining the collaborative filtering recommendation system's efficacy. To ensure the best use of storage and processing power, the last step is to assess how well resources have been allocated.

4.1. Dataset description. Cloud computing is popular because it allows users to pay-per-use for on-demand computing. Energy-aware work scheduling improves resource usage and is cost-effective. Traditional task, resource, and energy scheduling strategies fail with cloud computing. The main objective of the analysis is to compare hybrid and conventional cloud computing scheduling strategies. It compares Shortest work First (SJF), Round Robin (RR), Max-Min, and Min-Min work scheduling algorithms [21]. The best resource scheduling algorithms are STAR, Dynamic Resource Allocation Scheme, Autonomous Agent-Based Load Balancing, Credit-Based Scheduling Algorithm, Greedy-Based Job Scheduling Algorithm, Optimal Algorithm, Resource-Aware Hybrid Scheduling Algorithm, and Honeybee Algorithm. Energy-based scheduling approaches for carbon-reducing scheduling are investigated last. The analysis found that Min-Min surpasses the competition in turnaround and waiting time.

4.2. Analysis of scalability of distributed storage:. The work aims to assess the scalability of CC-DRAM's distributed storage technology is described in Fig.4.1. Performance under varying loads and content

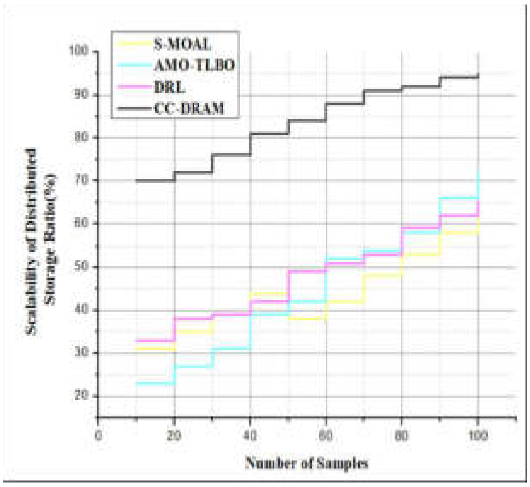


Fig. 4.1: The Graph of Scalability of Distributed Storage

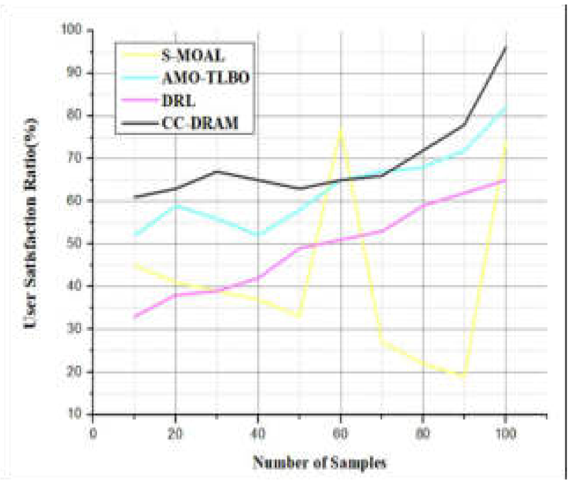


Fig. 4.2: The Graphical Representation of User satisfaction

sizes evaluates the system's capacity to handle increasing user demand and large volumes of instructional materials. According to the results, the distributed storage system can scale well, meaning that it can handle more users and more data without sacrificing speed or availability. Because of its adaptability, CC-DRAM can suit a wide range of educational needs without compromising on economy or speed. The distributed storage scalability is analysed in this proposed method and the values are obtained by the ratio of 97.8% is shown in Fig.4.1.

4.3. Analysis of user satisfaction. The purpose of the analysis is to determine how happy CC-DRAM platform users are with the service generally and with the personalized recommendations in particular which is shown in Fig.4.2. Teacher and student questionnaires and feedback forms show widespread approval of the proposed course of analysis, particularly with regard to its practicality and applicability. The importance of customized suggestions in improving educational outcomes is underscored by user reports of higher engagement and a more tailored learning experience. Positive comments about CC-DRAM indicate that it meets the needs of its target audience. In Fig.4.2, analysis of user satisfaction is improved in the proposed method by the ratio

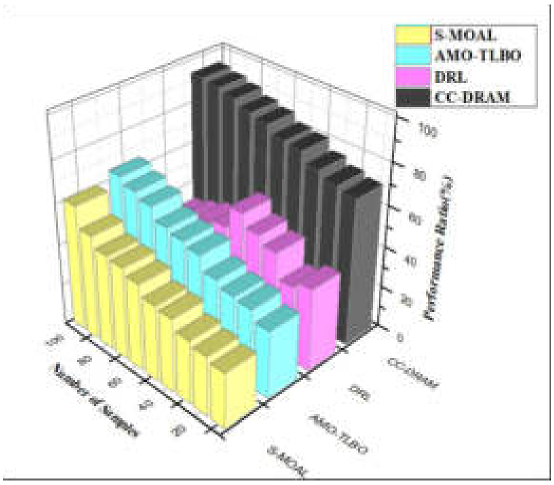


Fig. 4.3: The Graphical Illustration of Performance Ratio

of 98.2% compared to the existing method.

The entire performance of CC-DRAM is examined in this analysis, with a focus on system responsiveness, resource utilization, and uptime (figure 8). Cloud task scheduling and resource allocation efficacy is measured by metrics including error rates, throughput, and latency. When compared with alternative memory technologies, the findings show that CC-DRAM continually provides superior performance, including short latency and efficient use of resources. In addition, the robust performance measurements show that CC-DRAM can handle the demands of online learning environments, guaranteeing that instructional contents can be accessed quick and reliably. The analysis of performance ratio is 99.34% which is increased in the proposed method is shown in Fig.4.3.

Examining how well CC-DRAM's collaborative filtering recommendation system works is the main goal of the present analysis is shown in Fig.4.4. Analyzing data from user interactions and comments, the analysis determines how well the algorithm can understand user preferences and provide accurate and relevant content suggestions. The results show that the collaborative filtering approach effectively makes the learning experience more personalized, with users being more engaged and the content being more relevant. Keeping students' attention and enhancing educational outcomes both depend on its effectiveness. In this proposed method the analysis of effectiveness of collaborative is achieved by the ratio of 96.12% is shown in Fig.4.4.

4.4. Analysis of resource allocation. The analysis aims to assess how well and efficiently CC-DRAM allocates its resources which is explained in the Fig.4.5. The efficiency with which the system allocates storage and processing power in response to user needs and activity patterns is examined. It was found that dynamic resource allocation ensures efficient use of resources, which leads to less waste and ensures that instructional materials are delivered effectively. By distributing information strategically, one can save students from being overwhelmed with data, guarantee that they will only get engaging and relevant materials, and improve their education as a whole. Compared to existing method the analysis of resource allocation is increased by the ratio of 98.41% in this proposed method.

This paper assesses the CC-DRAM platform on many metrics, including its capacity to scale, user happiness, performance, collaborative filtering effectiveness and resource allocation efficiency. The analysis shows that the distributed storage system can readily scale to handle more users and data without lowering performance or availability standards. People are very happy with it, especially when it comes to personalized recommendations. Better utilization is shown by using overall performance metrics while considering aid use and responsiveness. Collaborative filtering permits for a greater tailor-made mastering enjoy, which boosts hobby and retention way to greater applicable content material. Allocating sources well prevents loss and makes sure that teaching gear are used successfully, which in the end results in a higher studying experience normal. When as compared to

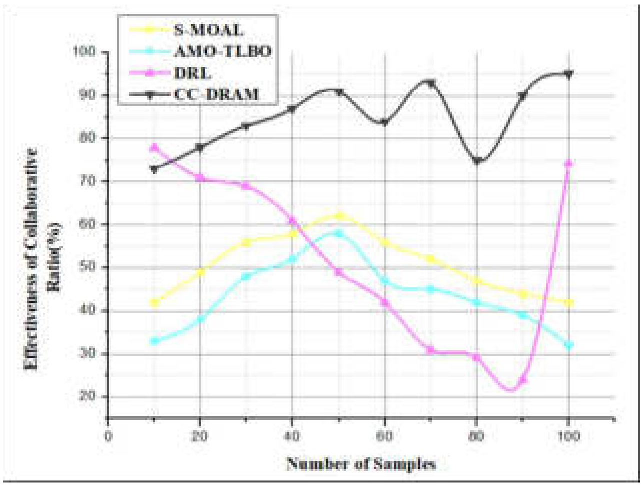


Fig. 4.4: The Graph of Effectiveness of Collaborative

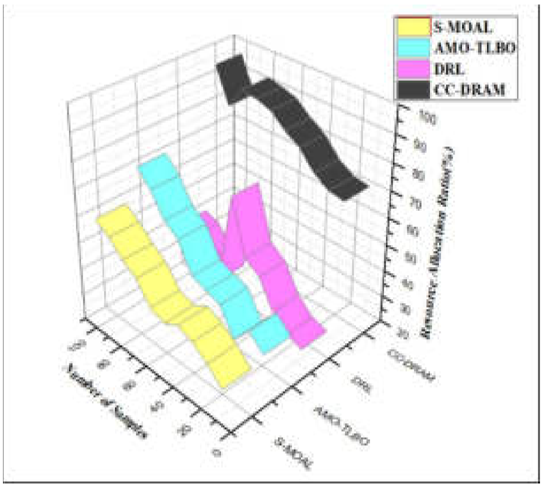


Fig. 4.5: The Graphical Representation of Resource Allocation

the modern-day procedures, every unmarried analysis indicates good sized improvements.

5. Conclusion. The CC-DRAM, which is founded on Cloud Computing, is accountable for a good sized enhancement that complements the skills of on-line studying platforms. This progress is carried out thru the availability of powerful resource management and educational stories that are individualized to the scholar. It is possible for CC-DRAM to personalize the dissemination of material to every person pupil via the usage of a collaborative filtering notion method. This ensures that scholars get assets that are tailor-made to their specific regions of interest. Enhancing mastering results while fending off the downsides of conventional online school rooms, which includes information overload, is one of the advantages of this technique.

5.1. Future work. Future studies will focus on improving the accuracy and adaptableness of advice algorithms as its key number one goal. With the assistance of modern-day device studying techniques, it could additionally be feasible to attain improved customization and predictive analytics. Our primary objective is to enhance the model with extra capabilities that can be interesting to apply and academic assets that can be beneficial. As a result, we believe that CC-DRAM can adjust to the ever-evolving requirements of educators in the modern-day virtual environment.

5.2. Funding.

1. Teaching Reform Project of Colleges and Universities in Hunan Province ”Construction and Practice of Online and Offline First-class Course of International Settlement Based on OBE Concept”, Project No.: HNJG-20231111
2. Project of social science achievement evaluation committee of Hunan Province in 2024:”Research on the promotion of ideological and political education in the era of artificial intelligence” Project No. XSP24YBC033

REFERENCES

[1] De la Prieta, F., Rodríguez-González, S., Chamoso, P., Demazeau, Y.,& Corchado, J. M. (2020). An intelligent approach to allocating resources within an agent-based cloud computing platform. Applied Sciences, 10(12), 4361.

[2] Qiu, C., & Ding, F. (2021, October). Research on Applied Undergraduate Education Resource Allocation System Based on Cloud Computing. In 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA) (pp. 800-802). IEEE.

[3] Shang, Q. (2021). A dynamic resource allocation algorithm in cloud computing based on workflow and resource clustering. Journal of Internet Technology, 22(2), 403-411.

[4] Yang, F. (2021). CLOUD COMPUTING VIRTUAL RESOURCE DYNAMIC SYSTEM ALLOCATION AND APPLICATION BASED ON SYSTEM ARCHITECTURE. Dynamic Systems and Applications, 30(5), 753-770.

[5] Sohani, M., & Jain, S. C. (2021). A predictive priority-based dynamic resource provisioning scheme with load balancing in heterogeneous cloud computing. IEEE access, 9, 62653-62664.

[6] Wu, W., & Plakhtii, A. (2021). E-learning based on cloud computing. International Journal of Emerging Technologies in Learning (IJET), 16(10), 4-17.

[7] Shukur, H., Zeebaree, S., Zebari, R., Zeebaree, D., Ahmed, O., & Salih, A. (2020). Cloud computing virtualization of resources allocation for distributed systems. Journal of Applied Science and Technology Trends, 1(2), 98-105.

[8] Maithili, K., Vinothkumar, V., & Latha, P. (2018). Analyzing the security mechanisms to prevent unauthorized access in cloud and network security. Journal of Computational and Theoretical Nanoscience, 15(6-7), 2059-2063.

[9] Panwar, R., & Supriya, M. (2022). Dynamic resource provisioning for service-based cloud applications: A Bayesian learning approach. Journal of Parallel and Distributed Computing, 168, 90-107.

[10] Khan, S., Al-Dmour, A., Bali, V., Rabbani, M. R., & Thirunavukkarasu, K. (2021). Cloud computing based futuristic educational model for virtual learning. Journal of Statistics and Management Systems, 24(2), 357-385.

[11] Karthiban, K., & Raj, J. S. (2020). An efficient green computing fair resource allocation in cloud computing using modified deep reinforcement learning algorithm. Soft Computing, 24(19), 14933-14942.

[12] Wei, J., & Mo, L. (2020). Open interactive education algorithm based on cloud computing and big data. International Journal of Internet Protocol Technology, 13(3), 151-157.

[13] Kumar, Y., Kaul, S., & Hu, Y. C. (2022). Machine learning for energy-resource allocation, workflow scheduling and live migration in cloud computing: State-of-the-art survey. Sustainable Computing: Informatics and Systems, 36, 100780.

[14] A. C. Kaladevi, V. Vinoth Kumar, A. K. Velmurugan, K. Gunasekaran, B. Swapna and V. Dhilip Kumar, ”Realization and Prediction of IoT-based Dynamic Social Interactions for the Future Recommendations”, Ad Hoc & Sensor Wireless Networks, 58.3-4, p. 243-271.

[15] Fan, L., Xia, M., Huang, P., & Hu, J. (2021). Research on educational information platform based on cloud computing. Security and Communication Networks, 2021, 1-11.

[16] Belgacem, A., Beghdad-Bey, K., Nacer, H., & Bouznad, S. (2020). Efficient dynamic resource allocation method for cloud computing environment. Cluster Computing, 23(4), 2871-2889.

[17] Moazeni, A., Khorsand, R., & Ramezanpour, M. (2023). Dynamic resource allocation using an adaptive multi-objective teaching-learning based optimization algorithm in cloud. IEEE Access, 11, 23407-23419.

[18] Venkatesan, V. K., Ramakrishna, M. T., Batyuk, A., Barna, A., & Havrysh, B. (2023). High-Performance Artificial Intelligence Recommendation of Quality Research Papers Using Effective Collaborative Approach. Systems, 11(2), 81.

[19] Deng, X., Zhang, J., Zhang, H., & Jiang, P. (2022). Deep-reinforcement-learning-based resource allocation for cloud gaming via edge computing. IEEE Internet of Things Journal, 10(6), 5364-5377.

[20] Shi, F., & Lin, J. (2022). Virtual machine resource allocation optimization in cloud computing based on multiobjective genetic algorithm. Computational Intelligence and Neuroscience, 2022.

[21] Data search: https://datasetsearch.research.google.com/search?src=0& query=Cloud%20compbased%20dynamic%20resource %20allocation%20Model& docid=L2cvMTF5M2.

[22] Karthikeyan, T., Sekaran, K., Ranjith, D., & Balajee, J. M. (2019). Personalized content extraction and text classification using effective web scraping techniques. International Journal of Web Portals (IJWP), 11(2), 41-52.

[23] Kumar, V. V., Muthukumaran, V., Ashwini, N., Beschi, I. S., Gunasekaran, K., & Niveditha, V. R. (2022). An efficient signcryption scheme using near-ring hybrid approach for an IoT-based system. International Journal of e-Collaboration (IJeC), 18(1), 1-31.

[24] Anbazhagu, U. V., Niveditha, V. R., Bhat, C. R., Mahesh, T. R., & Swapna, B. (2024). High-performance technique for

item recommendation in social networks using multiview clustering. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL, 19(1)..

[25] Rajalakshmi, V., Muthukumaran, V., Koti, M. S., Vinothkumar, V., & Thillaiarasu, N. (2022). E-Collaboration for Management Information Systems Using Deep Learning Technique. In Handbook of Research on Technologies and Systems for E-Collaboration During Global Crises (pp. 398-411). IGI Global..

[26] Khasawneh, N. A. S., Khasawneh, A., Khasawneh, M. A. S., & Jadallah abed Khasawneh, Y. (2024). Improving Arabic Content Delivery on Cloud Computing Platforms for Jordanian E-learning Environments. Migration Letters, 21(S1), 575-585.

[27] Medichalam, K. ., Vijayarajan, V. ., Kumar, V. V. ., Iyer, I. M. ., Vanukuri, Y. K. ., Prasath, V. B. S. ., & Swapna, B. . (2023). Trustworthy Artificial Intelligence and Automatic Morse Code Based Communication Recognition with Eye Tracking. Journal of Mobile Multimedia, 19(06), 1439–1462. https://doi.org/10.13052/jmm1550-4646.1964

[28] Naveed, Q. N., Qahmash, A. I., Qureshi, M. R. N., Ahmad, N., Abdul Rasheed, M. A., & Akhtaruzzaman, M. (2023). Analyzing critical success factors for sustainable cloud-based mobile learning (CBML) in crisp and fuzzy environment. Sustainability, 15(2), 1017.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jul 7, 2024
Accepted: Sep 14, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1132–1146, DOI 10.12694/scpe.v26i3.4268

DEVELOPMENT OF DEEP LEARNING-BASED MEDIA CONTENT RECOMMENDATION SYSTEM, DL-MCRS, FOR USER SATISFACTION

QI LUO[*]

Abstract. The ever-growing quantity of audio-visual information accessible today may be effectively managed by recommender systems, which assist users in discovering new and interesting topics. An increasing number of customized suggestion apps have emerged on the World Wide Web in the last decade. Recommendation systems can't function without precise behaviour modelling of users. The conventional wisdom about friend suggestion algorithms leaves out crucial user data, leading to a misleading portrayal of their actions. The common understanding of friend suggestions is inaccurate because it disregards crucial user information. Hence, this paper proposes that the Deep Learning-Based Media Content Recommendation System (DL-MCRS) improves efficiency and user satisfaction by integrating huge multi-source heterogeneity data and building more precise user and item models on social media platforms. The suggested method uses the semantic personalized recommendation system (SPRS) to bridge the gap between high-level semantic information and low-level media properties. The suggested system uses domain ontology to customise video recommendations to their interests based on a user's past actions on the site. The experimental findings show that the suggested strategy outperforms the baseline methods concerning efficiency.

Key words: Deep Learning; Social Media; User; Recommendation System.

1. Introduction. Many limitations hinder the development of DL-MCRS that would improve customer happiness, which are systems suggesting media content using deep learning [1]. A key challenge is the need for large volumes of diverse, unique data to train effective deep learning models [2]. Getting such data is difficult and often accompanied by regulatory obstacles that respect consumer privacy as well as data protection [3]. Another big obstacle is ensuring transparency and understandability in deep learning models [4]. Additionally, recommendation system must adapt to changing user tastes in real time [5]. This requires a lot of computer power and complex algorithms to ensure it remains accurate and relevant [6]. Biased recommendations can negatively affect user experience and perpetuate stereotypes; hence, the biases within training datasets need to be addressed too [7]. To enhance customer participation and prevent echo chambers there should be a trade-off between personalized rules and accidental contents discovery[8]. Scalability, the ability to handle increasing numbers of users and content with no compromise on performance, is another important factor that needs consideration[9]. Finally, there are methodological and technical hurdles to clean when incorporating person feedback into the gadget to decorate pointers over the years [10]. Data management, model transparency, computational efficiency, bias mitigation, and user interaction are important components that ought to be carefully considered for DL-MCRS to acquire its full potential in enhancing person happiness via personalised content suggestions [11].

Improved customer delight has been a riding pressure in the back of the speedy evolution of DL-MCRS [12]. Methods inclusive of content material-based filtering, collaborative filtering, and hybrid methods are critical [13]. Both user-primarily based and object-based totally collaborative filtering leverage statistics from user interactions to predict what users will primarily based on their shared choices [14]. To offer hints which might be similar to what a user has loved, content material-based filtering takes into account the inherent characteristics of media content material, such genre, actors, or keywords [15]. To get around each technique's weaknesses, hybrid approaches integrate them to provide higher, extra tailor-made answers. With the proliferation of online resources like video streaming sites, social media, and news organizations, people's media consumption has skyrocketed in the modern digital age. Despite the vast amount of information, consumers often experience frustration and decreased engagement due to their inability to locate media that suits their interests. There are

[*] School of Journalism and Communication, Hunan Mass Media Vocational Technical College, Changsha, Hunan, 410100, China.

a lot of problems with the current recommendation systems, which are rule-based or depend on collaborative filtering. They provide recommendations that are either irrelevant or repetitious since they do not understand consumers' subtle preferences. Beyond that, these algorithms may not be able to adjust to the ever-shifting media environment or consumers' ever-changing preferences.

Acquiring complex styles in consumer behaviour and content attributes has been performed through the use of deep getting to know models RNNs and CNNs [16]. Some examples of CNN and RNN capabilities encompass CNN's ability to interpret visual content functions from movies and snap shots and RNN's superiority in managing sequential facts, which makes them well-acceptable to comprehending consumer interaction sequences across time. To similarly enhance advice accuracy, autoencoders and Variational autoencoders (VAEs) are employed to accumulate compact, latent representations of both users and objects.

Several boundaries nonetheless remain within the development of DL-MCRS, even with these advances. A major impediment is the bloodless-begin problem, which makes it hard to make suitable pointers for brand new customers or products with little interplay information. Since deep getting to know images necessitate effective computing assets and effective algorithms to manage big quantities of information, their scalability is another situation. Improving patron happiness fully requires ongoing innovation, the integration of varied statistics sources, and superior modelling tools to cope with those problems.

- By combining information from many sources, the suggested DL-MCRS tackles the problem of handling the abundance of audio-visual data. With this integration, people can learn more about the user's habits and preferences, which improves the quality of our recommendations.
- The DL-MCRS uses SPRS to fix the semantic gap between media attributes and high-level semantic information. Because of this, the system can improve user happiness by making tailored suggestions that are highly relevant to each user's interests and preferences.
- Experimental validation shows that the proposed DL-MCRS is more efficient than baseline approaches. Improved user satisfaction is the end result of the system's optimisation of resource utilisation and suggestion accuracy through the customisation of video recommendations based on domain ontology and user interactions.

Developing a Media Content Recommendation System (DL-MCRS) that Uses Deep Learning to Make Users Comfortable. Section III deals with the results of the DL-MCRS, a media content recommendation system that is based on deep learning. Section IV presents the findings and analysis, Section V follows with a discussion, and ends with a summary and some recommendations.

2. Literature Survey. As recommendation systems field expands, several approaches have come up to address the issue of personalized suggestions for media consumption. DL-MCRS happens to be one of such choices. The review gives an overview of current research done in DL-MCRS hence attention on how they have improved recommendation accuracy & user satisfaction.

Sharma et al. [17] make use of semantic personalised recommendation system (SPRS) which utilizes domain ontology together with user activity so as to recommend videos. Performance is measured using predicted ratings compared against actual ratings showing that there was an improvement in precision, recall, accuracy over standard metadata-based systems. This SLR conducted by Da'u , A. et al. [18] focuses on recommender systems(RSs) based on Deep Learning: including some useful findings by other authors regarding this topic . Autoencoder models are mostly used followed by CNNs and RNNs. The most popular datasets are MovieLens, as well as Amazon reviews. Measures of evaluation in this area include precision, RMSE. Da'u , A. et al., [19] use a tensor factorization (TF) machine for overall rating prediction, which implements aspect-based opinion mining (ABOM) with a multichannel deep convolutional neural network (MCNN) for aspect extraction as well as aspect-specific rating generation. When compared to the baseline approaches, the results reveal notable improved accuracy. Deep autoencoders are used by Shambour, Q. [20] to capture complex user-item associations to improve the accuracy of a multi-criteria recommender system (M-CRS). According to experiments carried out on Yahoo! Movies and TripAdvisor datasets, the algorithm outperforms state-of-the-art recommendation engines by producing more precise predictions. Khanal, S. S. et al., [21] give an overview of recommendation systems in e-learning(RS-E-L), which are classified into content-based, knowledge-based, collaborative filtering or hybrid systems. Components such as algorithms from machine learning methodologies; datasets; evaluation techniques represent some key areas presented under the taxonomy.

Weijin Di [22] suggested constructing a personalized learning content Recommendation system based on a recommendation algorithm in English learning (CPLRS-EL-RA). At first, the Movielens-1M dataset is used for data collection. The next step is to begin pre-processing using the acquired data. The Generalized Moment Kalman Filter, or GMKF, is a tool used to pre-process data. The pre-processing output is fed into the feature extraction process using the Enhanced Synchro Extracting Wavelet Transform (ESWT) to extract the students' attitudes, connections, and entities. In the next step, the recommendation algorithm is given the extracted output. Listening, speaking, reading, and writing are the four areas of learning that the recommendation algorithm successfully categorizes. Using the Tiger Beetle Optimizer (TBO), the weight parameter of the Recommendation Algorithm may be optimized. Utilizing criteria such as accuracy, precision, recall, sensitivity, specificity, and calculation time, the efficiency of the suggested technique is evaluated after its activation in Python. When compared to other methods, such as PRSETR-CRNN, LCBCRS-CNN, and HRSC-ANN, the CPLRSEL-RA method achieves better accuracy (22.32%), sensitivity (27.32%), and recall (31.13%), sensitivity (24.43%), and recall (38.13%), respectively, for listening.

Manikandan and Kavitha [23] proposed the Harris Hawks Optimization, Cuckoo search and Deep Semantic Structure Model (DSSM) for content recommendation systems for e-learning. New optimization algorithms like the Enhanced Personalized Best Cuckoo Search Algorithm (EpBestCSA) and the Enhanced Harris Hawks Optimization Algorithm (EHHOA) are used in the suggested content recommendation system's semantic-aware hybrid feature optimizer to choose appropriate features that improve prediction accuracy. Another new algorithm, the DSSM, combines Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). The suggested model beats competing recommendation systems regarding accuracy, recall, f-measure, and prediction precision, as shown in the experiments. The suggested technique is tested using ten-fold cross-validation.

Balakumar Muniandi et al. [24] recommended the Deep Learning Approach for Adaptive Content Recommendation Systems for Digital Marketing Platforms. Focusing on its application utilizing deep learning methods, this study investigates the function of adaptive content recommendation systems in digital marketing platforms. The paper explores the fundamentals, approaches, and difficulties of creating and implementing such systems. The author demonstrates the efficacy of deep learning methods in improving the accuracy of content recommendations and user engagement by doing a thorough literature and case study evaluation. The author discusses developments and improvements in the future.

Arnav Dubey et al. [25] presented the Digital Content Recommendation System through Facial Emotion Recognition. The paper's opening section introduces and briefly discusses various face emotion detection algorithms. A summary of the literature on these algorithms pertaining to music and movie recommendation systems follows. The paper's second section includes a discussion of the possible advantages of incorporating face expression detection into music and movie recommendation systems. Two of these are possible improvements to the user experience and the capacity to provide more tailored suggestions depending on the user's emotional condition. The paper's third section presents a look at the pros and cons of using face expression detection in music and movie recommendation systems. Concerns around privacy and ethical problems are among them, as are challenges with the algorithms' accuracy and trustworthiness.

Wenhua Liu [26] introduced the digital entertainment content recommendation algorithm for user behaviour analysis of an English learning social platform. This article learns about learners' preferences, habits, and past learning by collecting and analyzing their data, building learning models, and other relevant information. Afterwards, a collaborative filtering algorithm is employed to match students with peers who exhibit comparable interest preferences and learning styles, classify students into similar groups, make recommendations based on the choices and preferences of similar students, and ultimately provide students with the best learning materials. This article builds a full suite of social capabilities to encourage learners to connect and share information and provide individualized recommendations. The findings demonstrated that the proposed social platform for learning English was feasible and beneficial through study and testing with real user data. According to user feedback, learning results and communication between learners have been greatly enhanced by the platform's social features and individualized recommendations.

This summary helps understand ongoing studies and identify issues related to the subject matter. DL-MCRS's research findings prove its better performance over other content recommendation mechanisms making it an ultimate leader in this area when one needs personalized recommendations.

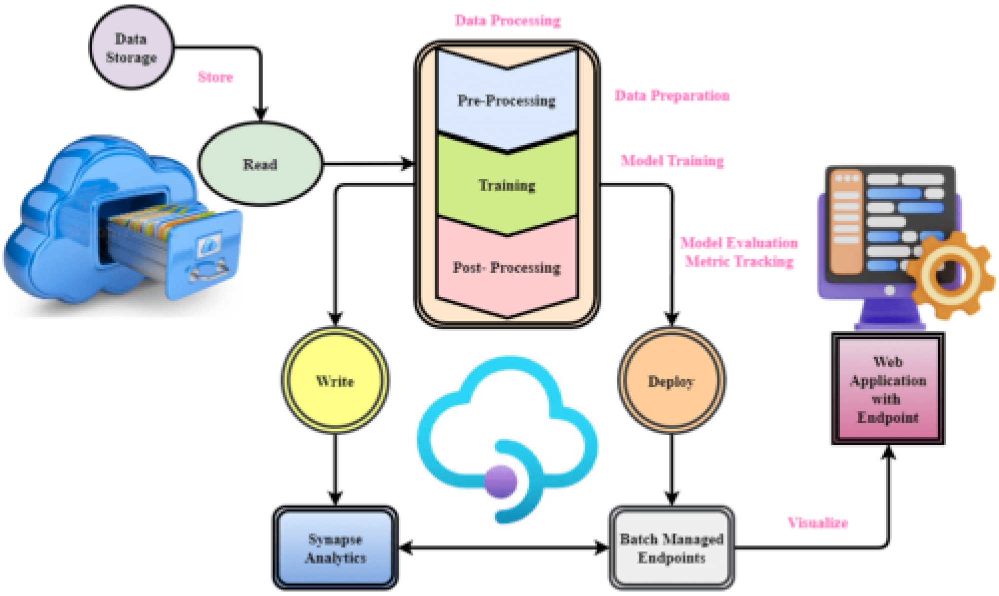


Fig. 3.1: Schematic of a content-based recommendation system

3. Proposed Method. Users wanting interesting and pertinent media have a tremendous hurdle in the digital era due to the enormous number of audiovisual information available online. Subpar user experiences are a common result of traditional recommendation algorithms' inability to properly understand user behavior. In this paper, present DL-MCRS, which uses heterogeneous multi-source data to improve efficiency and customer satisfaction. Incorporating a SPRS allows DL-MCRS to successfully synchronize high-level semantics with a small amount of media property. Given a user's past behaviors on social media, this novel method uses a domain ontology to provide personalized video suggestions.

A content-based suggestion system, like the one shown in the picture, uses artificial intelligence to provide unique suggestions for each user. In Figure 3.1, there's data storage, which is essentially a repository for all the pertinent information, including user profiles, item properties, and user-item interactions. Data Processing follows, which includes operations like cleansing, separating features, and conversion to get the data ready for training the model.

Algorithms learn correlations and trends between users and things based on content attributes through Model Training, which uses pre-processed data. The Model Assessment Metric Tracking is concerned with the performance of these models in terms of accuracy, recall and precision. These models are produced and then used on a real-time basis to provide recommendations based on customer inputs. While it remains an important part of this process, Synapse Analytics is an integrated analytics platform that stores, processes and deploys models. It supports scalable data processing for seamless integration of multiple data sources necessary for wide-ranging recommendation systems. The flow diagram helps stakeholders grasp the process behind building content-based recommendation system; from data collection to model deployment where we want end users to get individual suggestions which are effective as well as efficient.

$$Q_k^{(2)} = [Q_{k1}^{(2)}, Q_{k2}^{(2)}, ...Q_{k(\frac{n-i_1+1}{2})}^{(2)}] + (m_1 - m_2) \tag{3.1}$$

By modifying its components according to the disparity between two criteria, $m_1 - m_2$, Equation (3.1) depicts the change of a vector $Q_k^{(2)}$. To be more precise, each member of the adjusted latent parameter vector for user,

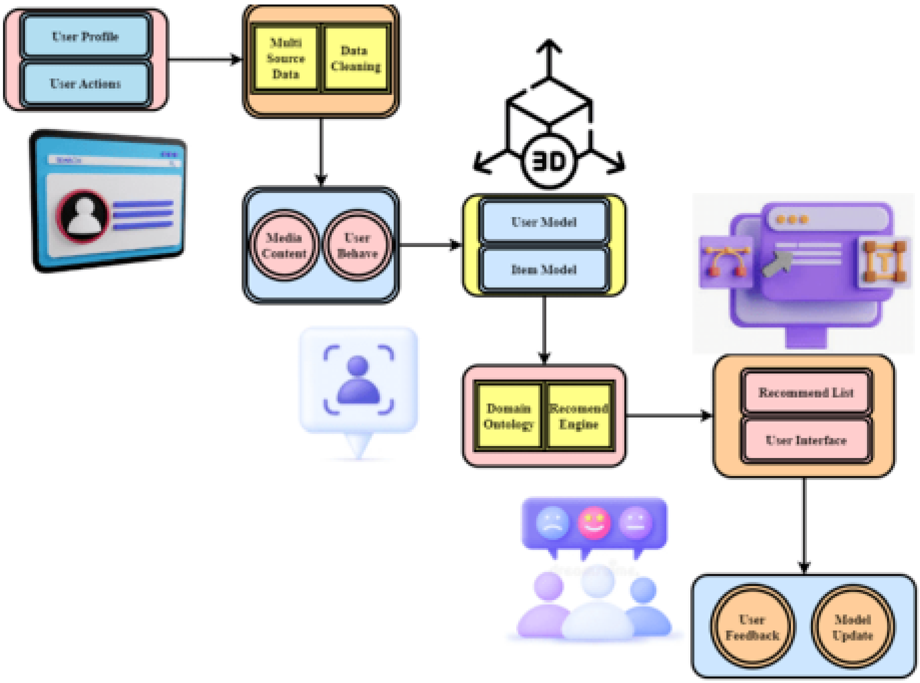


Fig. 3.2: Design of content-based system for recommendations

where $Q^{(2)}_{k1}, Q^{(2)}_{k2}, ...Q^{(2)}_{k(\frac{n-i_1+1}{2})}$.

$$d^{(u)}_{kj} = h\big(g_h \times Q^{(u-1)}_{k,j:j+h}\big) + c_u(e+q) + \big(w_q(1+p)\big) \tag{3.2}$$

For a user k at a certain layer u in a deep learning model, Equation (3.2) describes the calculation of a feature $d^{(u)}_{kj}$. Here, the convolution process that was done to the latent vector of features Q from the preceding layer $u-1$, capturing specific trends in user behavior, is denoted as $g_h \times Q^{(u-1)}_{k,j:j+h}$. The non-linear transformation is applied through the function h. The model may be fine-tuned using the extra parameters and bias terms introduced by $c_u(e+q)$ and $w_q(1+p)$.

$$Q^{(p)}_v = \big[Q^{(u)}_{k2}, Q_{pu}(u), ...Q_{w(\frac{n-z_y+1}{2})}\big] - w_q(1+p) \tag{3.3}$$

The process of creating a vector $Q^{(p)}_v$ by integrating different feature parts and modifying them with a bias term is illustrated by Equation (3.3). The aggregated vector of features for a person or object at a certain stage is represented by $Q^{(u)}_{k2}$ The vector contains components that include multi-level features obtained from the deep learning layers, such as $Q_{pu}(u)$ and $Q_{w(\frac{n-z_y+1}{2})}$, among others. To enhance the feature vector and effectively represent consumers' nuanced preferences, a regularization factor is introduced by $w_q(1+p)$.

Figure 3.2 shows the architecture of content-based recommendation system showing several levels and processes involved in providing personalized recommendations to the users by Service providers. Beginning at the User Interface Layer, user profiles and actions are collected and recorded here. Between the computer system and its users, it allows information sharing and feedback channels to be open. In addition to all this, the layer gets data from user actions and processes that involve cleaning up and integrating different sources qualitatively for better quality assurance. It finally cleanses the data further before moving it into another phase of analysis.

Feature extraction efforts of Feature Extraction layer focus media and user actions primarily. User behavior analysis emanates from records over interactions like preference or trends while media content analysis extracts contents like text as well as photos including metadata about them. Such aspects give their features to deep-learning Models whose categories include Item attributes/User preferences understanding models among others within them . Eventually, these models use those characteristics they have retrieved as input so that they may make personalized suggestions. A Semantic Customized Recommendations System uses domain ontology along with recommendation engines to generate personal recommendations based on individual preferences also ensuring that suggested items are most notionally similar ones with those that were chosen by a user.

$$T_k = \frac{f^{wk}}{\sum_{l=1}^{U} f_x w} + (w+q) \sim \frac{(qp+rt)}{e} d_s^z(p+q) \tag{3.4}$$

Equation (3.4) describes the process of calculating a parameter T_k that incorporates several weighted features together with a regularization factor. To ensure proportionate effect, the weighted characteristic T_k is normalized over the combined amount of all characteristic weights by adding them together. Adding $(w+q) \sim \frac{(qp+rt)}{e}$ creates a complicated adjustment factor that combines recurring user activity $d_s^z(p+q)$ with features of the media.

$$S_{v,j} = s_p + \omega \sum_{W \neq P}^{W} cos(v,z) + (f_{pk} - s_t) \mp (d+e) \tag{3.5}$$

Equation (3.5) shows how to calculate a score $S_{v,j}$ by taking into account different factors that impact the suggestion. The relevance score utilized for user content ranking in the (DL-MCRS) is denoted by s_p. The base score is added by $cos(v,z)$, and the weighted cosine similarity measurements between vectors $(f_{pk} - s_t)$ are integrated by $d+e$, which captures similarities in customer preferences and item characteristics.

$$s_{v,k} = F(s_{l,m}) = \sum_{l=0}^{P} j \times Qs_{(d_{j+k})} = g^{q+k} \tag{3.6}$$

The calculation of a score $s_(v,k)$ for a user and a specific characteristic is illustrated in Equation (3.6), which relies on an aggregate function F applied to intermediary scores $s_(l,m)$. The last user feature score utilized to tailor content suggestions is denoted by $Qs_((d_(j+k)))$ in (DL-MCRS). Over a variety of characteristics, with weights j, the summation $g^{(q+k)}$ aggregates scores with weights.

Profiles of users and suggestion content learning: In this stage, want to develop a model that is individual to each user so that may anticipate their interest in (multimedia) goods by analysing their past interactions with them. To provide content suggestions that are specific to the preferences of the target user, the learnt user profile structure is compared to item profiles, which reflect representative item attributes. Video recordings are a kind of media that is complicated. Many facts are communicated to us (by the writer) through many multimedia channels, especially the visual and auditory ones, while watch a film. Because are based on human-generated data and are believed to cover the information meaning of movies to a large extent, most movie recommendation systems currently use content-based filtering (CBF) or collaborative filtering (CF) models. These models rely on metadata, such as genre or the wisdom of the crowd, to make recommendations. Learning from multi-modal inputs (e.g., audio, visual, and metadata) and the data collected from multimedia material, on the other hand, might help us comprehend natural events in videos better by revealing links between different modalities.

The system begins with raw material from a video. The first step is to break the video down into its constituent frames. There is a single picture in the video sequence for every frame. It further divides each frame into smaller sections. To analyze visual material with more precision, these blocks are used as the building blocks for feature extraction. Vector representations that capture audio or visual input properties are called i-vectors. I-Vectors can condense the material into a compact representation. Features at the Block Level contain details like colours, textures, and other low-level visual attributes retrieved from each frame's frames. AlexNet, a CNN, is used to get in-depth features from the video clips. Examples of higher-level, abstract content features are item identification and scene comprehension. Aesthetic Visual Features (AVFs) evaluate the content's visual attractiveness by looking at colour harmony, picture composition, and other aesthetic aspects that could impact user choice. Summing up the characteristics extracted throughout time

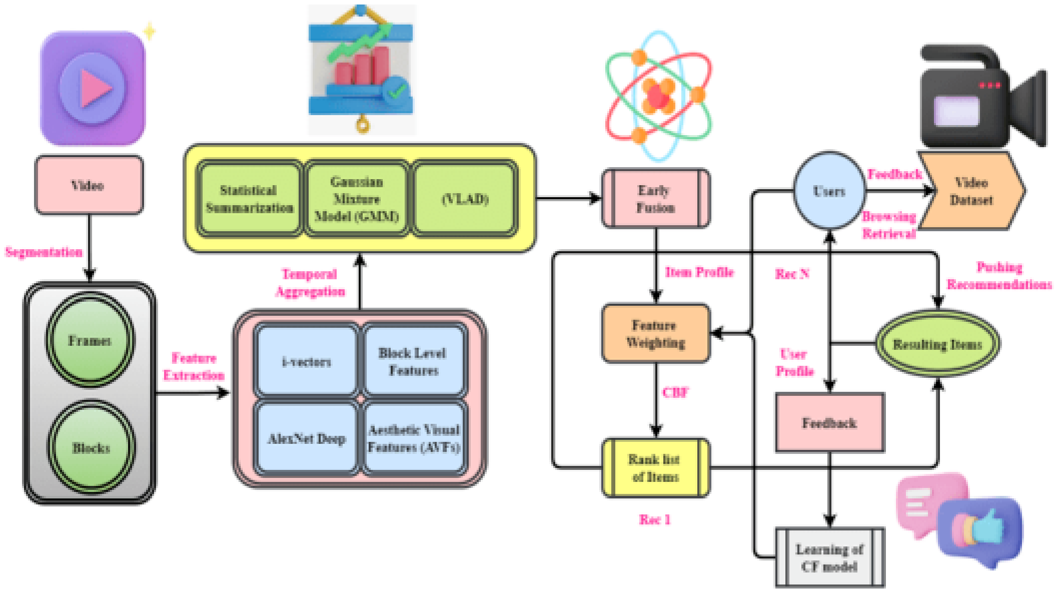


Fig. 3.3: Content based firltering collaboration

allows statistical summarization to capture the video's temporal dynamics. Some possible statistical metrics that depict the evolution of traits over time include variance, mean, and standard deviation. One probabilistic model that may be used to depict the distribution of video characteristics is GMM. Utilized mostly for content-based clustering and pattern detection, it can model the intricate distribution of data. The Vector of Locally Aggregated Descriptors aims to combine SIFT descriptors and other locally pertinent features into a single, fixed-length vector. It may use this vector representation, which is more compact to represent the whole video or parts of it. After temporal aggregation, i-Vectors, block-level features, deep features, and AVFs are fused. An item profile that better represents the video content is the result of this fusion, which uses the best features of each kind. An item profile, a thorough video description, is created from the combined characteristics. In this profile, you can find all the attributes that have been retrieved and aggregated. These features will be used in the recommendation process. The content and the user's choices determine the relative relevance of other characteristics. It may fine-tune the recommendation process by using a weighting mechanism to prioritize some attributes over others. To provide content-based filtering, the system makes use of the item profile. In contrast to collaborative filtering, which utilizes data on user interactions, content-based filtering (CBF) makes recommendations based on the specifics of the material itself. The system gets a ranked list of items (Rec 1, Rec N) from the weighted features and content-based filtering. According to the characteristics of the material, these are the things that the user would find most interesting. Customers engage with the platform by perusing the suggested products and offering comments (likes, dislikes, viewing, skipping). This connection is vital to improving the recommendation process. If the system gets enough input, it can train and refine its collaborative filtering model. CF's recommendation system is based on user activity trends, such as how often others with similar preferences use it. The model considers this feedback loop to improve its ability to foretell users' preferences.

These unique traits of diverse feature sets cater to the varying information requirements of consumers in figure 3.3.

$$rsf(s,p) = \frac{\sum_{j<k}^{p}(s_{j,p} - u_r) \times (f_{p,k} - e_d)}{\sqrt{\alpha} j \forall_{k,p} + (s_{p,i} - e_f)^2 + s} \tag{3.7}$$

The computation of a suggested score functional $rsf(s,p)$, which integrates different elements to provide an ultimate score, is defined by Equation (3.7). An item's relevance to a user is measured by $s_(j,p) - u_r$ within the framework (DL-MCRS). The total of all the weighted disparities between user preferences $(f_(p,k) - e_d)$ and an average user rating and between item features $\sqrt{\alpha} j \forall_{k,p}$ and a feature bias $s_(p.i) - e_f$ is captured.

$$tan(\partial, p) = \frac{\sum_{k}^{e}[f_{g+1}(p_w q_t)] + (W_q - z_{kl})}{\sqrt{\sum_{v \ni A}^{Q-P}(P_{v,j} - j_p)}} \tag{3.8}$$

The tangent operation, defined by Equation (3.8) as $tan(\partial, p)$, is used to calculate a score that is dependent on several user and object feature components. A transformation that is used to improve the accuracy of recommendations might be represented by $f_(g+1)(p_w q_t)$. The weighted features and biases are aggregated in the numerator $W_q - z_k l$, which captures the complicated interplay between various user preferences and item characteristics $P_{v,j} - j_p$.

$$sec(p,q) = \frac{v.pq}{(q) + 1(pq)} = \frac{\sum_{j+w}^{(st+p)}(s_{f+z})(e_s)}{\sqrt[2]{\sum_{j \times \forall}^{p}(q+1)}} \tag{3.9}$$

To improve the suggestions by taking into consideration certain user-item interactions, $sec(p,q)$ might be defined in Equation (3.9). To ensure proportionality, the weighed product is normalized by a factor of adjustment and the first half $\frac{v.pq}{(q)+1(pq)}$ is represented. The second portion, which is the total of weighted scores $(s_(f+z))(e_s)$, divided by a normalization term containing the sum of adjusted interactions, is expressed as $\sum_{j \times \forall}^{p}(q+1)$

One solution to the problem of information overload is a recommendation system, which uses the user's interests, preferences, or past actions to sift through massive amounts of dynamically created content and extract meaningful pieces of information. Put another way, a system for recommendations can utilize a user's profile to determine the likelihood that would enjoy a certain item. These recommendation algorithms are useful for both businesses and consumers. When it comes to buying things online, these systems help customers save money on things like searching for information, choosing products, and making a final decision. Consequently, recommendation algorithms have become widely used on e-commerce platforms.

Helping people make better decisions through individualized recommendations increases their happiness. There are essentially two types of approaches used in recommendation systems to examine user preferences (Figure 3.4). One is filtering based on content, which uses product features like related keywords to narrow search results. One kind of product suggestion is the content-based (CB) method, which looks at the user's past purchases to determine what other items would enjoy. The basic idea behind content-based recommendation systems is as follows: first, take a look at what a user likes, figure out what features share, and add those preferences to their profile. Then, compare those features to the user's profile and suggest products that are very similar.

$$tan_{(u+1)} + (q, mp) = \frac{w}{q} + (e_{s+p}) + (p,q) \tag{3.10}$$

The sum of many weighted components is represented by the converted score $tan_((u+1)) + (q, mp)$, which is computed using Equation (3.10). By dividing a weight w/q, the term ensures that the influence is balanced. A bias term derived from item and user attributes is introduced by adding $e_(s+p)$.

$$P_{(w+f)} + (u,t) = \sum_{j=0}^{p} qw_{(j+p)} + (x,p) \tag{3.11}$$

For the purpose of accuracy analysis, Equation (3.11) describes the weighted average $P_{(w+f)} + (u,t)$ of characteristics. The sum of weighted features $qw_{(j+p)}$ and a bias term x, p across a given range is represented by this equation.

$$\max_{(j,k)} = \frac{S_{mkl} \times (k + wp)(w - 2q)}{n_{=0,1,2}} + (p.q) \tag{3.12}$$

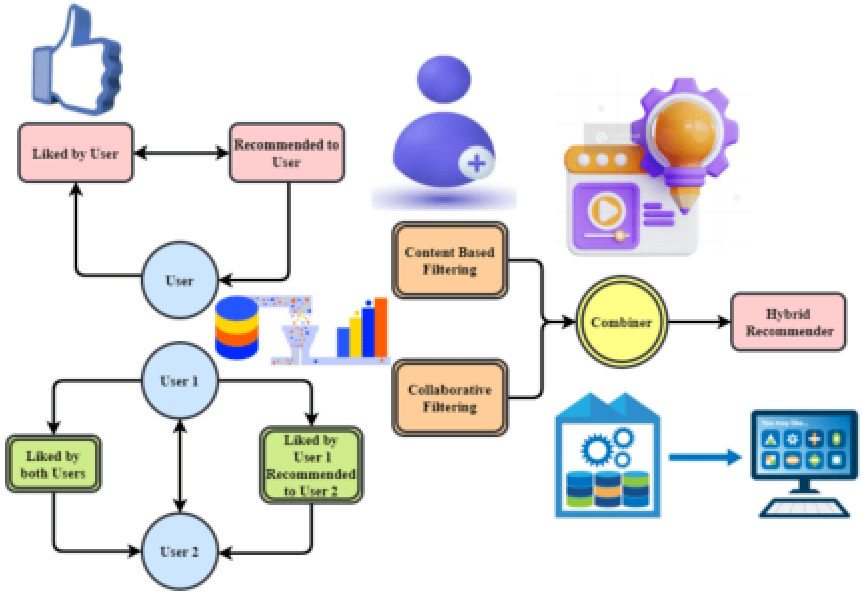


Fig. 3.4: Different types of filtering: content driven and cooperative

In user satisfaction analysis, Equation (3.12) is used to determine the greatest value $max_{(j,k)}$. Adjustments depending on item characteristics and system settings are introduced by $(k+wp)(w-2q)$, while $S_m kl$ signifies a satisfaction metric that captures user input or engagement. A normalization factor that incorporates different degrees of user interaction might be represented as $n_{(} = 0,1,2)$ when divided by this. In addition, the satisfaction measure may be fine-tuned by considering p, q as an interaction element between the parameters.

The overarching idea of RS is to leverage past interactions with material to determine and predict which items users will find interesting. Figure 3.5 shows the overall design of a standard RS. Likes, clicks, and ratings are examples of implicit feedback that users give the system when interacting with it. For example, if a user gives a new smartphone a high rating, she could be interesting in reading more articles on apps for mobile devices. Thus, using this information to deduce user interests is the fundamental notion of RS. The RS learns a model to anticipate the user's potential interest in new things based on their previous responses. As a further step, it rank the items based on how relevant to think will be to the user. At last, the user will be aggressively offered the things that rank higher. Depending on the domain of application, the ratings' semantic meaning might vary greatly. For example, most OSN employ binary values, whereas e-commerce websites and services that offer video on demand generally use discrete sets of ordered integers (like a 5-point rating scale).

$$z_{uk} = g\big(v, j|Q_p, w_{T-u}\big) = Q_k^{ku-p} \tag{3.13}$$

In the context of analysing computational efficiency the equation (13) explains a computational process $z_u k$ that is governed by a function g. In this context, v and j are probably indices for users and items, respectively, while Q_p and w_{T-u} stand for certain feature vectors or matrices. The output Q_k^{ku-p}, which may stand for a latent feature.

$$z_{pu} = h(p,k) = d_f^{r+1}\big(w + p(1+q)\big) \tag{3.14}$$

The scalable functional $z_p u$ described by Equation (3.14) is used in scalability analysis and is defined by h. At this point, it is probable that p, k stand for indices or variables, d_f^{r+1} signifies a feature vector magnified. A phrase that might stand for a versatile adjustment factor is the weighted sum of the system parameters,

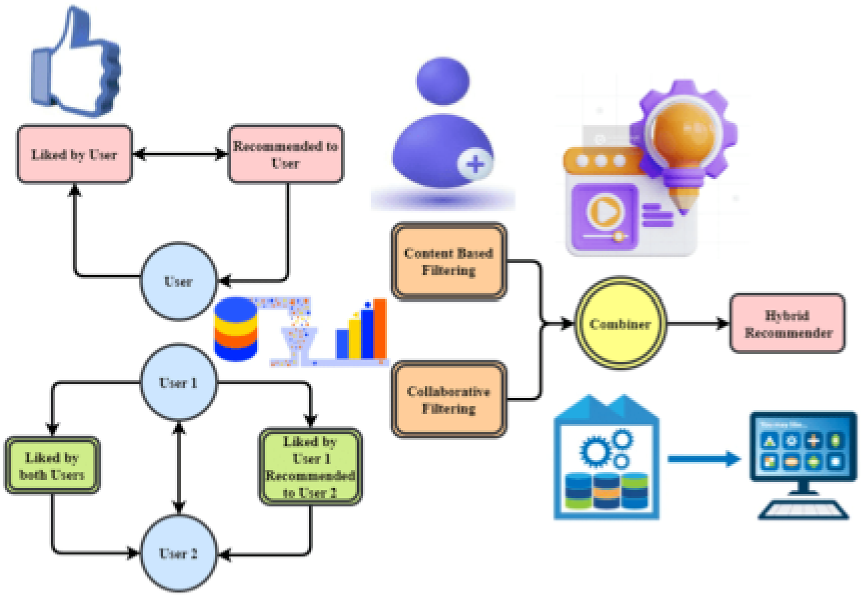


Fig. 3.5: Architecture of recommender system

denoted as w+p(1+q).

$$M = - \sum_{(v,j)\ni ZUZ}^{P} \left[f_g p + (hj) F_{e+s} \right] + (Q_l P) \tag{3.15}$$

In personalization analysis, Equation (3.15) is used to describe a summing process M. It is possible that ZUZ represents individualized interactions between users and items, whereas v,j probably stand for user and item indices. Characteristics of items and user preferences are probably represented by $f_g p, hj$, and F_{e+s}.

By improving upon previous approaches, the suggested DL-MCRS brings recommendations technologies a long way. To provide more accurate and tailored material suggestions, DL-MCRS makes use of complex user and item mathematical models in conjunction with large amounts of data from several sources. To further improve user happiness, (SPRS) can be included to guarantee a smooth relationship between high-level semantic and media attributes. The efficacy of DL-MCRS is proven to be higher than baseline approaches in experiments, which confirms its potential to improve the user interface and completely transform the way users receive content suggestions.

4. Results and Discussion. The performance of the proposed DL-MCRS has been analyzed based on metrics accuracy, user satisfaction, computational efficiency, scalability, and personalized analysis compared to conventional recommendation methods such as SLR-RS [18], ABOM [19], and M-CRS [20].

In Figure 4.1, analysing the DL-MCRS accuracy in delivering applicable content to users is the principle emphasis of accuracy analysis for user pleasure. The capability of DL-MCRS to faithfully simulate user movements and forecast user alternatives is vital to its efficacy. Accuracy evaluates how properly the device can perceive applicable objects, while Recall examines how well it can recognise all applicable gadgets. For a nicely-rounded assessment of accuracy, the F1-Score a harmonic mean of Precision and Recall is beneficial. Using MAE, you will study the everyday discrepancy among predicted and actual consumer scores produces 99.4 percentage. The DL-MCRS is capable of draw close elaborate styles in consumer interactions and media content as it makes use of today's deep mastering images like CNNs and RNNs. Autoencoders and Variational Autoencoders (VAEs) improve recommendation accuracy by using getting to know efficient latent representations of customers and objects. By effectively integrating multi-source heterogeneous data, DL-MCRS produces extra

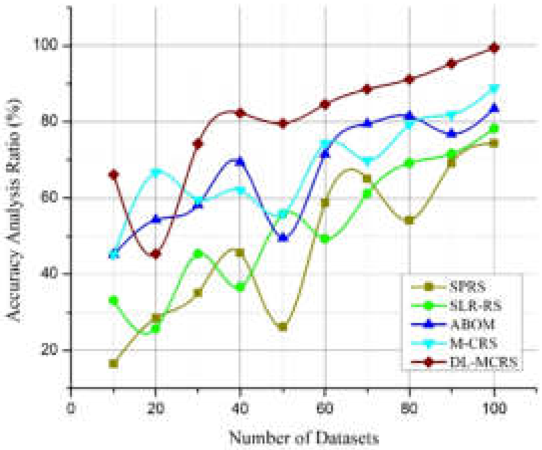


Fig. 4.1: Accuracy analysis

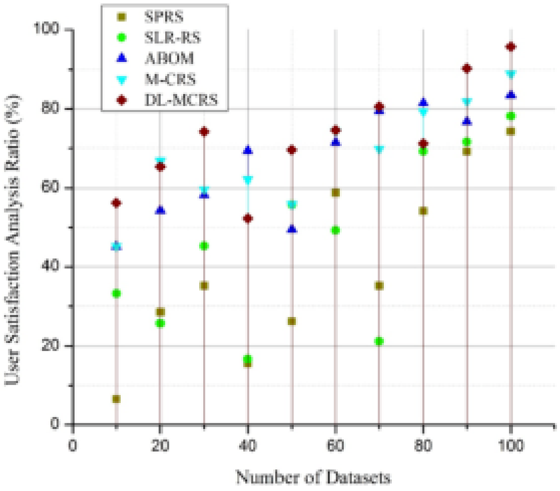


Fig. 4.2: User satisfaction analysis

accurate consumer and item images, leading to experimental evaluations that display its superior performance as compared to conventional advice systems. In addition, DL-MCRS's Semantic Personalised Recommendation System (SPRS) lets in for extra unique tips which are in keeping with consumer diversions via connecting high-degree semantic records with low-stage media attributes. With its multi-faceted accuracy analysis, DL-MCRS proves to be technically superior and has the capability to significantly enhance person happiness with the aid of handing over personalized content with pinpoint accuracy.

An evaluation of the DL-MCRS potential to satisfy its users' needs and choices is conducted as part of its person delight take an explore. Some essential indicators to recollect while assessing person happiness are engagement, reside length, click-via costs, and remarks rankings. In Figure 4.2, users find the encouraged cloth relevant and attractive if there's excessive user involvement and expanded live time. DL-MCRS uses state-of-the-art deep studying models to generate specific and tailor-made hints, which provides to user happiness. Combining CNNs and RNNs permits in-intensity exam of person moves and content material homes, guaranteeing that hints are noticeably relevant to user alternatives produces 95.7 percentage. Customers are

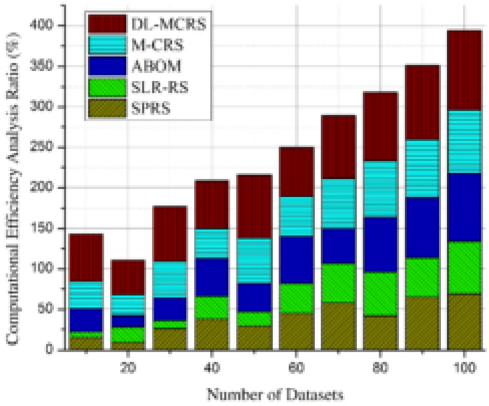


Fig. 4.3: Computational efficiency analysis

more satisfied due to the device's capacity to resolve frequent problems, like the cold-start problem, which takes place while there isn't sufficient data for proper hints for brand new users or matters. With the help of domain ontology and person records, DL-MCRS is able to unexpectedly alter to new users and gadgets even as maintaining the advice excellent true. According to empirical studies, DL-MCRS notably increases user pleasure metrics and exceeds traditional advice systems in terms of accuracy. With the device's capacity to offer entertaining and relevant media suggestions, users are more inclined to keep looking round and find out extra content.

In Figure 4.3, an efficient DL-MCRS is being advanced with the intention of increasing consumer satisfaction. Gathering records, cleaning it up, generating features, schooling the model, arising with tips, and subsequently evaluating it are all essential parts. Many measures are used to quantify computational performance, which includes schooling time, inference time, energy utilization, scalability, and aid utilisation. Tests towards both classic and modern-day images, as well as profiling equipment like TensorBoard and NVIDIA's Nsight Systems, permit for an assessment of overall performance. For example, at the same time as working with the MovieLens dataset, applying a Neural Collaborative Filtering (NCF) model requires optimising hyperparameters, tracking training periods, and creating low-latency inference pipelines. It is possible to identify bottlenecks by means of tracking resource usage and scalability below one of a kind masses. There is likewise an effort to reduce strength usage via investigating strategies like model quantization and trimming produces 98.3 percentage. The effects display how computational performance and recommendation first-class aren't mutually distinctive, and how optimised images appreciably boom both real-time overall performance and person happiness.

To guarantee consumer pride thru green and personalized recommendations at scale, it's far necessary to conduct a scalability evaluation whilst growing a DL-MCRS. In Figure 4.4, this necessitates monitoring the gadget's responsiveness to growing statistics masses and person counts. Capabilities for green model schooling, real-time inference performance, and statistics control are vital components. Data guidance techniques that are scalable can handle growing datasets without notably reducing overall performance. To determine how nicely a model can scale at some point of education, people study how a whole lot time and electricity it takes to use disbursed schooling methods and optimised hardware like GPUs and TPUs to train on bigger datasets. With the usage of strategies like caching, sharding, and model distillation, real-time inference scalability keeps advice latency low regardless of the number of users requesting the provider. Methods which includes vertical scaling, which makes use of more powerful hardware, and horizontal scaling, which entails adding greater processing nodes, are taken into consideration. These methods, while paired with ongoing monitoring and optimisation, assure that DL-MCRS can keep up its great overall performance and responsiveness, making customers happier whilst demand for will increase produces 94.5 percentage. To nicely manipulate the ever-increasing facts and consumer base, destiny improvement will centre on similarly optimising present scalable structures and learning emerging technology.

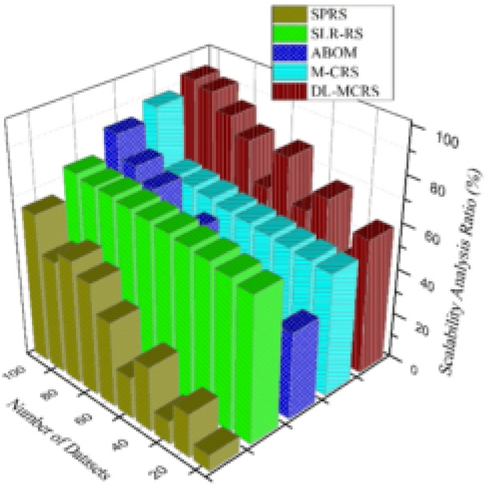


Fig. 4.4: Scalability analysis

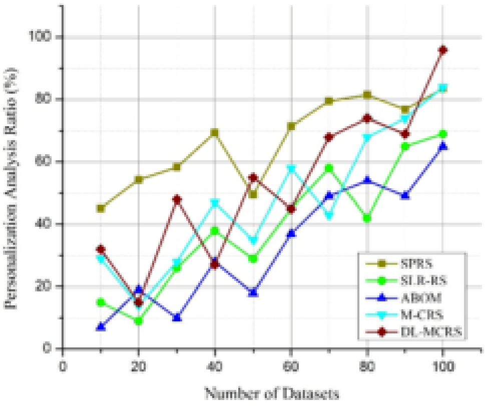


Fig. 4.5: Personalized analysis

In Figure 4.5, the creation of a DL-MCRS that aims to maximise user happiness relies heavily on personalisation analysis. This evaluation looks at how well the system uses complex algorithms to understand user behaviour and preferences to personalise content. The three main parts are selecting a model, extracting features, and user profiling. The process of user profiling entails collecting extensive information about user interactions, including ratings, watching history, and implicit feedback. Collaborative filtering, content-based filtering, or hybrid techniques are all examples of deep learning models that might benefit from feature extraction, which converts this raw data into useful inputs. To gauge how effectively the suggested material satisfies the user's interests, performance indicators including precision, recall, and mean reciprocal rank (MRR) are employed. To improve the accuracy of recommendations, advanced DL models recurrent neural networks and neural collaborative filtering record intricate interactions between users and items produces 96.7 percentage. By integrating real-time feedback and regularly retraining models, continuous learning processes are put in place to respond to developing consumer preferences. The optimal trade-off between customisation and computing efficiency is then optimised through experimental comparison of various models and methodologies. The study takes user diversity into account as well, making sure the system gives fair recommendations to all user categories. In the end, people want our recommendation algorithms to be as efficient and scalable as possible while still delivering highly relevant content that makes users happy and engaged. To further enhance personalisation capabilities, future work will investigate new deep learning architectures and integrate multi-modal data.

As a result, the paper's extensive research further establishes DL-MCRS as an important facilitator of individualised content distribution across different domains, demonstrating its superiority over conventional recommendation algorithms.

5. Conclusion. The suggested DL-MCRS solves a number of the problems with older recommendation structures. The DL-MCRS algorithm improves advice accuracy and person happiness via constructing greater correct and specific models of user behaviour and item attributes the usage of big-scale, multi-source heterogeneous statistics. Notable among these functions is the incorporation of a Semantic Personalised Recommendation System (SPRS), which correctly connects low-degree media residences to excessive-level semantic data. A greater personalised viewing revel in is made possible through the use of area ontology, which similarly improves the recommendation technique with the aid of coordinating video guidelines with user interests in line with their beyond interactions. The experimental results display that DL-MCRS is more efficient in processing and making content tips than baseline procedures, proving its superiority. The device's potential to manipulate sophisticated consumer statistics and media cloth highlights its capacity to revolutionise social media discovery and engagement. In addition to improving user happiness, DL-MCRS creates an extra dynamic and attractive media consumption environment via fixing problems the bloodless-start trouble and ensuring various content material is exposed. This method gives a stable foundation for similarly advancements in tailor-made media content material distribution and is therefore a first-rate soar ahead inside the records of recommendation systems.

REFERENCES

[1] Wen, X. (2021). Using deep learning approach and IoT architecture to build the intelligent music recommendation system. *Soft Computing,* 25(4), 3087-3096.

[2] Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., & Basilico, J. (2021). Deep learning for recommender systems: A Netflix case study. *AI Magazine,* 42(3), 7-18.

[3] Nassar, N., Jafar, A., & Rahhal, Y. (2020). A novel deep multi-criteria collaborative filtering model for recommendation system. *Knowledge-Based Systems*, 187, 104811.

[4] Huang, L., Fu, M., Li, F., Qu, H., Liu, Y., & Chen, W. (2021). A deep reinforcement learning based long-term recommender system. *Knowledge-Based Systems*, 213, 106706.

[5] Huang, H., Mu, J., Gong, N. Z., Li, Q., Liu, B., & Xu, M. (2021). Data poisoning attacks to deep learning based recommender systems. *arXiv preprint arXiv:2101.02644.*

[6] Tahmasebi, H., Ravanmehr, R., & Mohamadrezaei, R. (2021). Social movie recommender system based on deep autoencoder network using Twitter data. *Neural Computing and Applications,* 33(5), 1607-1623.

[7] Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex & Intelligent Systems,* 6(3), 621-634.

[8] Chiu, M. C., Huang, J. H., Gupta, S., & Akman, G. (2021). Developing a personalized recommendation system in a smart product service system based on unsupervised learning model. *Computers in Industry,* 128, 103421.

[9] Deldjoo, Y., Schedl, M., Cremonesi, P., & Pasi, G. (2020). Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR),* 53(5), 1-38.

[10] Venkatesan, V. K., Ramakrishna, M. T., Batyuk, A., Barna, A., & Havrysh, B. (2023). High-Performance Artificial Intelligence Recommendation of Quality Research Papers Using Effective Collaborative Approach. Systems, 11(2), 81.

[11] Afsar, M. M., Crump, T., & Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys,* 55(7), 1-38.

[12] Koloveas, P., Chantzios, T., Alevizopoulou, S., Skiadopoulos, S., & Tryfonopoulos, C. (2021). intime: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence. *Electronics,* 10(7), 818.

[13] Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. *Procedia Computer Science*, 167, 2318-2327.

[14] Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning. *arXiv preprint* arXiv:2403.19345.

[15] Ahmed, S. T., Kumar, V. V., & Kim, J. (2023). AITel: eHealth augmented-intelligence-based telemedicine resource recommendation framework for IoT devices in smart cities. IEEE Internet of Things Journal, 10(21), 18461-18468.

[16] Dang, C. N., Moreno-García, M. N., & Prieta, F. D. L. (2021). An approach to integrating sentiment analysis into recommender systems. *Sensors*, 21(16), 5666.

[17] Sharma, S., Rana, V., & Kumar, V. (2021). Deep learning based semantic personalized recommendation system. *International Journal of Information Management Data Insights*, 1(2), 100028.

[18] 18) TR, M., Vinoth Kumar, V., & Lim, S. J. (2023). UsCoTc: Improved Collaborative Filtering (CFL) recommendation

methodology using user confidence, time context with impact factors for performance enhancement. PLoS One, 18(3), e0282904.

[19] Da'u, A., Salim, N., Rabiu, I., & Osman, A. (2020). Recommendation system exploiting aspect-based opinion mining with deep learning method. *Information Sciences,* 512, 1279-1292.

[20] 20) Anbazhagu, U. V., Niveditha, V. R., Bhat, C. R., Mahesh, T. R., & Swapna, B. (2024). High-performance technique for item recommendation in social networks using multiview clustering. INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL, 19(1).

[21] Khanal, S. S., Prasad, P. W. C., Alsadoon, A., & Maag, A. (2020). A systematic review: machine learning based recommendation systems for *e-learning. Education and Information Technologies*, 25(4), 2635-2664.

[22] Ahmed, S. T., Kumar, V. V., Singh, K. K., Singh, A., Muthukumaran, V., & Gupta, D. (2022). 6G enabled federated learning for secure IoMT resource recommendation and propagation analysis. Computers and Electrical Engineering, 102, 108210.

[23] Manikandan, N. K., & Kavitha, M. (2023). A content recommendation system for e-learning using enhanced Harris Hawks Optimization, Cuckoo search and DSSM. Journal of Intelligent & Fuzzy Systems, 44(5), 7305-7318.

[24] Ramakrishna, M. T., Venkatesan, V. K., Bhardwaj, R., Bhatia, S., Rahmani, M. K. I., Lashari, S. A., & Alabdali, A. M. (2023). HCoF: Hybrid Collaborative Filtering Using Social and Semantic Suggestions for Friend Recommendation. Electronics, 12(6), 1365.

[25] Dubey, A., Shingala, B., Panara, J. R., Desai, K., & Sahana, M. P. (2023). Digital Content Recommendation System through Facial Emotion Recognition. Int. J. Res. Appl. Sci. Eng. Technol, 11, 1272-1276.

[26] Ahmed, S. T., Vinoth Kumar, V., Mahesh, T. R., Narasimha Prasad, L. V., Velmurugan, A. K., Muthukumaran, V., & Niveditha, V. R. (2024). FedOPT: federated learning-based heterogeneous resource recommendation and optimization for edge computing. Soft Computing, 1-12.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jul 9, 2024
Accepted: Sep 19, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1147–1164, DOI 10.12694/scpe.v26i3.4300

DEEP LEARNING BASED STACKED PROBABILISTIC ATTENTION NEURAL NETWORK FOR THE PREDICTION OF BIO MARKERS IN NON-HODGKIN LYMPHOMA

SIVARANJINI NAGARAJAN[*]AND GOMATHI MUTHUSAMY[†]

Abstract. The biomolecular characterization of Non-Hodgkin lymphoma (NHL) impacts the prognosis, therapy planning, and prediction of therapeutic response. The development of cancerous characteristics in lymphoma formation may often be attributed to certain genetic defects and the resulting disruption of oncogenic regulatory processes. The use of advanced technology has made it feasible to identify genetic variations and their corresponding biomarkers. However, the current challenges in histopathology include the identification techniques and the presence of different cell types inside a tumour. Computational techniques are now being used more often to diagnose genetic abnormalities without invasive procedures. This is done by analysing quantitative imaging data. Therefore, we are now deploying a deep learning-based stacking probabilistic attention neural network in this project. In this study, the histopathological images are obtained from the Kaggle source. Next, the image may undergo preprocessing using the soft switch Weiner filter (SSWF). The area of interest was segmented using the hierarchical seed polarity transform (HSPT). The biomarker linked with Non-Hodgkin lymphoma is categorised using the stacked probabilistic attention neural network (SPANN) based on the segmented output. The whole experiment was conducted using a histopathologic cancer dataset from Kaggle under python environment. The proposed strategy outperformed the current state-of-the-art alternatives by obtaining high range of accuracy(95%), precision(95%), recall(95%) and F score (92%).

Key words: Non-hodgkin lymphoma, Bio marker, deep learning, soft switch weiner filter, hierarchical seed polarity transform , stacked probabilistic attention neural network

1. Introduction. B-non-Hodgkin lymphomas (B-NHLs) are a subgroup of B-cell lymphomas that often display characteristics resembling the first phases of normal B-cell maturation. Flow cytometry, immunohistochemistry (IHC), immunoglobulin clonality assessment, fluorescence in situ hybridization (FISH), and next-generation DNA sequencing may be employed together with standard cytogenetics to more precisely classify these cancers. Protein expression is assessed by doing immunohistochemical staining on tissue sections placed on glass slides. This data is then used to guide clinical decision-making in many diagnostic scenarios, including cancer classification, detection of remaining illness, and identification of mutations. Standard brightfield chromogenic Immunohistochemistry staining, when performed at a high-throughput level, has limitations such as a restricted range of variation and images that have a significant overlap between the chromogen and the stain. This requires the use of specialised digital techniques to separate and deconvolve the stains as a preprocessing step for advanced research and commercial quantification algorithms used in Immunohistochemistry. Additional research is necessary to find dependable biomarkers for NHL. Despite thorough hyper-parameter tuning on a case-by-case basis or the laborious and error-prone manual tagging of many markers linked with NHL, colour separation remains suboptimal in areas with significant chromogen overlap. Multiplex immunofluorescence (mpIF) staining is more effective than brightfield immunohistochemistry (IHC) staining because it enables the analysis of multiple markers either separately (without the need for stain deconvolution) or together (as a composite). This leads to enhanced co-localization, standardised staining, objective scoring, and determination of thresholds for all marker values, particularly in regions with low expression that are challenging to evaluate using IHC staining. A new meta-analysis suggests that deep learning has the potential to replace the labor-intensive manual detection methods presently employed for gene expression profiling or immunohistochemically

[*]Department of Computer Science, Auxilium College (Autonomous), Vellore, & Periyar University, Salem, Tamil Nadu, India (sathiya.siva5@gmail.com)

[†]Department of Computer Science,Government Arts and Science College, Komarapalayam, Tamil Nadu, India (mdgomathi@gmail.com)

stained photographs. These methods are costly and limited due to the paucity of multiplex immunofluorescence (mpIF) testing. By using computational tools, which provide several benefits, we have a unique chance to improve the prognosis of the most lethal illnesses. Although co-registered high-dimensional imaging of the same tissue samples can offer crucial reference data for the superimposed brightfield IHC channels, current deep learning methods depend exclusively on unreliable manual annotations, which suffer from unclear cell boundaries, overlapping cells, and difficulties in assessing low-expression regions. Our method utilises a unique deep learning technique, using a stacked probabilistic attention neural network, to achieve more precise categorization of biomarker cells with enhanced gene specificity. Using a single registered IHC and training data from the same slides enables this.

Our method utilises a unique deep learning technique, using a stacked probabilistic attention neural network, to achieve more precise categorization of biomarker cells with enhanced gene specificity. Using a single registered IHC and training data from the same slides enables this. A trained stacked probabilistic attention neural network can effectively detect NHL biomarkers using just an immunohistochemistry (IHC) picture as input.

This study aims to accomplish the following objectives.

- In-order to get the precise output soft switch weiner filter based preprocessing was used.
- To separate the region of interest from the image hierarchical seed polarity transform was used.
- For classifying the cancer associated biomarker stacked probabilistic attention neural network model was implemented.

The rest of the study is laid out as follows. We shall review the current research in this field in the second section. The statement of the issue is given in Section 3. Our methodology's outline may be found in Section 4. Section 5 describes our approach's implementation and assessment. The conclusion section of our analysis is in Section 6.

2. Related works. Lymphomas are malignancies that originate in certain cells of the immune system. They are categorised into two primary groups: Hodgkin lymphomas (HL) and non-Hodgkin lymphomas (NHL. HL and NHL vary in their growth patterns and microscopic appearance. The early identification of these diseases is essential because of its considerable influence on treatment results. Some of the strategies shown here have potential as future versions of NHL prediction systems.

The main objective of [1] was to emphasise the need of including predictive biomarkers. Initially, artificial intelligence (AI) was used to the data obtained from a specific dataset (GSE10846) including the gene expression profiles of 414 patients. A combination of machine learning and predictive analytics models, including the C5.0 algorithm, logistic regression, Bayesian Network, discriminant analysis, random trees, tree-AS, and Chi-square Automatic Inference, was employed to decrease the number of dimensions in the investigation of a potential relationship between overall survival and other clinicopathological variables.

The author of [2] conducts a morphologic analysis of histological sections from 209 patients with DLBCL, together with clinical and cytogenetic data. We used tissue microarrays (TMAs) made from three identical core slices to perform staining for CD10, BCL6, MUM1, BCL2, and MYC using H&E and immunohistochemical stains. The pathologists have assigned labels to the tissue microarrays (TMAs) indicating the regions of interest (ROIs) that specifically identify tissue samples that test positive for diffuse large B-cell lymphoma (DLBCL). We used a deep learning model to detect specific areas of interest (ROIs), isolate and classify all cancer cell nuclei inside those ROIs, and quantify various geometric properties for each nucleus. Gene expression analysis has shown its utility in predicting the success of DLBCL therapy [[3], [4]]. The author of [4] suggests a novel approach to enhance the selection of optimal disease targets for a multilayer biomedical network by using PPI data that is annotated with stable information from OMIM diseases and GO biological processes. The author presents enough evidence to substantiate the efficacy of the RecRWR approach.

The author of [5] uses two approaches, namely MIDER (Mutual Information Distance and Entropy Reduction) and PLSNET (Partial least square based feature selection), to analyse data and establish the topology of a Gene Regulatory Network (GRN) by computational means. Gene expression analysis were used to demonstrate both methodologies in the context of inflammatory bowel disease (IBD), pancreatic ductal adenocarcinoma (PDAC), and acute myeloid leukaemia (AML). All the genes that regulate these three pathways have been identified. The UGT1A gene family was shown to have a critical role in regulating inflammatory bowel illness in the dataset. Similarly, the SULF1 and THBS2 genes were discovered as important factors in the pancreatic cancer dataset. Furthermore, they demonstrate that combining the results of the MIDER and PLSNET methods may result in a more precise ensemble-based strategy for inferring the topology of the gene regulatory network from data. Furthermore, an approximate estimate for the sample size of upcoming validation tests was established. They proposed an analytical approach that may identify potential regulator genes for validation testing and determine the required sample size for these studies. The objective of the suggested augmented ensemble learning approach in [6] is to improve the speed and accuracy of medical diagnosis. This model has been used to investigate a diverse array of ailments, including Alzheimer's, pancreatic, brain, and breast malignancies. The results indicate that the proposed model surpasses the existing techniques in terms of both accuracy and latency.

The author of [7] provides a concise overview of the current state of research and clinical use of MRI biomarkers in cancer therapy. This article provides a comprehensive discussion of MRI biomarkers, including the method of collecting and preprocessing MRI data, as well as the use of machine learning techniques. It concludes with an overview of the many types of biomarkers and their clinical utility in various cancer types.

A method for categorising solid lung cancer that has been treated before, based on the detection of anaplastic lymphoma kinase (ALK) gene rearrangement, was established in [8]. Scientists at [9] aimed to develop a deep learning system capable of directly predicting the immunohistochemistry (IHC) phenotype using whole-slide images (WSIs). This would enable more precise subtyping of lung cancer using resected and biopsied tissues. The objective of the study [10] was to provide an automated method for quantifying CMYC. In order to determine the proportion of cancer cells that express CMYC, researchers use attention-based multiple instance learning. This method involves analysing tissue microarray cores that have been evaluated by a pathologist.

The author of [11] selected the expression of the Ki-67 protein as a molecular information proxy. The researchers proposed a deep convolutional network model to predict the presence of Ki-67 positive cells using H&E stained slides. The researchers gathered images of cells that were labelled as either negative or positive for Ki-67, along with pictures of the surrounding tissue and the microscope plate. These images were then used to train the algorithm. Slides that have been stained with haematoxylin and eosin may be analysed for follicular lymphoma (FL) using an innovative deep-learning algorithm. The programme's accuracy is determined by a confidence estimate level set in a previous study [12]. A Bayesian neural network (BNN) was trained, tested, and scored using whole-slide images of lymph nodes exhibiting FL or follicular hyperplasia.

The researcher in [13] used deep learning techniques to develop a software application capable of detecting the MYC rearrangement in digital histology slides of diffuse large B-cell lymphoma. Slides stained with hematoxylin and eosin (H&E) were used for the purpose of instructing and evaluating medical students and professors from a total of 11 distinct institutions.

The author of [14] created a multitask deep learning system named DeepLIIF to address the challenges of stain deconvolution/separation, cell segmentation, and quantitative single-cell IHC scoring simultaneously. This paper presents a new dataset that combines co-registered immunohistochemistry (IHC) and multiplex immunofluorescence (mpIF) staining on the same slides. We use this dataset to convert affordable IHC slides into more informative but costly mpIF images. Additionally, we utilise this dataset to provide the required reference information for the overlaid brightfield IHC channels. The author has devised a gene expression test [15] that can differentiate between the seven most prevalent subtypes of B-cell NHL. This study uses ligation-dependent reverse transcription polymerase chain reaction (RT-PCR) and next-generation sequencing to investigate the expression of more than 130 genetic markers. The main objective of the method was to restore microenvironmental indicators of gene expression linked to B-NHL cells. We used a random forest methodology for classification, which we trained and validated using a dataset of more than 400 cases exhibiting diverse histology. The therapeutic effectiveness of the treatment was shown by the restoration of cell-of-origin signatures and the normalisation of MYC and BCL2 expression levels in high-grade lymphomas. Additionally, the treatment successfully prevented major misclassification in low-grade lymphomas. Therefore, this highly accurate pan-B-NHL predictor, which allows for a methodical assessment of several diagnostic and prognostic indicators, may be suggested as a supplementary tool to conventional histology in guiding patient management and enhancing patient classification for pharmacological trials.

The author in [16] explores the capacity of machine learning (ML) techniques to enhance the Cox Proportional Hazard (CoxPH) model. The authors thoroughly analyse the flaws in the most recent version of the CoxPH model and then provide a diverse array of remedies, including both established and innovative approaches. The accuracy of the models is evaluated using two metrics: the Brier score and the concordance index. Ultimately, drawing on our discoveries, they provide a series of recommendations on how practitioners might effectively capitalise on the latest advancements in AI.

The paper [17] outlines a method for subtyping NHLs by combining transfer learning (TL) with principal component analysis (PCA).When implemented on disorganised data, the scalable approach described in [18]—a Neural network—produces dependable results.

The author of [19] developed a MUltiple SUV Threshold (MUST)-segmenter to identify tumours on PET scans. This method involves placing seed points and then extending them into areas.The study investigated the integration of clinical, molecular genotype, and radiomics characteristics in predicting the prognosis of individuals with aggressive B-cell lymphoma [20]. We used fluorescent in situ hybridization to examine gene rearrangements of MYC, BCL2, and BCL6.

3. Problem statement. As of from the literature survey the primary scientific challenge in oncology is to identify the tumours or genes responsible for cancer and their mutational interactions with other organ systems in the body. The primary challenge in analysing this data is its unstructured and varied morality. The data are sourced from several places, resulting in the following issues:

1. The data annotations that are not sequential throughout a wider array of patients.
2. The annotation fails to provide any insights into the therapeutic actions necessary to enhance data quality.
3. To expedite drug discovery for therapeutic development.
4. To expedite drug discovery for therapeutic development.
5. Timely selection of appropriate medication is essential due to the absence of longitudinal data about the survival duration of cancer patients within the community.
6. Fragmentation of clinical data across organisations, incompatibility of data standards, and lack of system interoperability result from inadequate methods for the diffusion of innovation.
7. The intelligent and effective storing of extensive gene expression or image data is very challenging.
8. Even a skilled scientist finds it hard to manually evaluate the data, since the motivation for using learning technologies, transferring, and obtaining vast amounts of data is very time-consuming. Consequently, researchers have recently used AI-based deep learning methods for precise prediction.

The use of histological evaluation of tissue sections at different levels of magnification has supplanted the reliance on morphological characteristics seen by haematoxylin and eosin (H&E) staining as the primary method for a pathologist to suspect the presence of lymphoma. Machine learning has gained popularity in cancer research due to its ability to extract complex information from medical pictures. Multiple radiomic characteristics are derived from pictures; nonetheless, machine learning necessitates appropriate parameters, therefore demanding meticulous feature selection.

Nevertheless, machine learning still encounters some challenges, including:

1. The precision of the model is influenced by the calibre of the photos used throughout the training process. The accuracy of the results may be compromised when using low-resolution photographs. Several variables, including as the scanner's precision, the uniformity of slide fabrication, and the quality of the stain, might potentially affect the picture.
2. The extensive variety of diseases, tissues, cells, and antibodies that are accessible suggests that it may be difficult to establish a direct correlation between morphological and molecular data. We are now concentrating on one specific connection, but more effort is needed to apply our technique more broadly.
3. The determination of whether portions of an H&E-stained image include positive or negative cells can only be made by referring to the matching IHC-stained picture. Despite using IHC staining, accurately determining the level of positivity of a cell in an H&E stained image remains challenging, hence impeding precise inference of the model.

Deep learning (DL) has been a powerful technology in the last decade since it can directly extract characteristics from photos. The area of computer vision has advanced as a consequence. DL models need vast amounts of input data because to the intricate nature of the underlying layers. When training a highly complex network with a limited dataset, the probability of overfitting is much higher. Techniques like as data augmentation,

Table 2.1: Comparative performance analysis

Disease	Ref.	Methodology	Remarks	Drawbacks
HL and NHL	[23]	IF and Machine learning	This technique facilitates the concurrent observation of various lymphoma cells, promotes computational learning and identification, and assists in discovering therapies and enhancing the knowledge of lymphoma.	High error rate
	[24]	Supervised machine	By using these several techniques together, they create a robust and intelligent computational instrument. This tool assists physicians in comprehending the potential impact of Hodgkin's lymphoma on individuals.	Low range of accuracy
	[27]	PET-CT and Ann Arbor	PET-CT scans and the Ann Arbour staging system are essential instruments for detecting and staging lymphoma, facilitating treatment choices. They provide precise staging for certain lymphoma types, categorising patients into phases and informing successful treatment approaches.	Not a cost effective one
HL and NHL	[26]	EACCED machine learning	SEER's cause-specific death categorisation is a valuable prognostic instrument; nevertheless, its efficacy is contingent upon data quality and needs continuous development and validation.	Conventional algorithm makes the process a time consuming one
LM and NLM	[29]	Digital image analysis	Digital image analysis and deep learning methodologies are transforming lymphoma diagnosis by automating histological investigation, discerning intricate patterns, and minimising subjectivity, hence enhancing patient outcomes.	High training rate
T-cell and B- cell Lympho mas	[28]	AI models using CNN	AI models are used for detecting DLBCL and addressing obstacles in imaging, data gathering, and privacy, demonstrating excellent diagnostic accuracy and the possibility for improved patient outcomes.	High training rate and cost expensive
	[25]	J48	The research used the J48 machine learning algorithm and the WEKA platform to construct diagnostic algorithms, which may enhance the accuracy of lymphoma categorisation, underscoring the significance of dependable tools in medical research.	Conventional algorithm makes the process a time consuming one
HL and NHL	[26]	EACCEED machine learning	SEER's cause-specific death categorisation is a valuable prognostic instrument; nevertheless, its efficacy is contingent upon data quality and needs continuous development and validation.	Conventional algorithm makes the process a time consuming one
LM and NLM	[29]	AI using CNN	Digital image analysis and deep learning methodologies are transforming lymphoma diagnosis by automating histological investigation, discerning intricate patterns, and minimising subjectivity, hence enhancing patient outcomes.	Time consuming process
T-cell and B- cell Lympho mas	[28]	J48 algorithm	AI models are used for the diagnosis of DLBCL, addressing obstacles in imaging, data collecting, and privacy, while achieving high diagnostic accuracy and the promise for improved patient outcomes.	Highly expensive and need hardware support
	[25]		The research used the J48 machine learning algorithm and the WEKA platform to construct diagnostic algorithms, possibly enhancing the accuracy of lymphoma categorisation and underscoring the need of dependable tools in medical research.	Unable to analyze the drawback because the result range was not given properly

transfer learning, and cross-validation may be used to address issues such as overfitting and insufficient data sets. Utilising cross-validation to forecast model uncertainty is a prevalent practice within the AI safety field. Moreover, it is crucial to interpret DL-based findings in order to provide comprehensible results for human assessment. This is essential for evaluating the safety of AI and expediting the integration of DL in practical medical contexts. The determination of positive or negative cells in portions of an H&E-stained picture can only be made by referring to the matching IHC-stained image. Despite using IHC staining, accurately determining the level of positivity of a cell in an H&E stained picture remains challenging, hence impeding more precise inference of the model.

Hence here in order to overcome all the existing issues we implement the deep learning based stacked probabilistic attention neural network for the prediction of NHL.

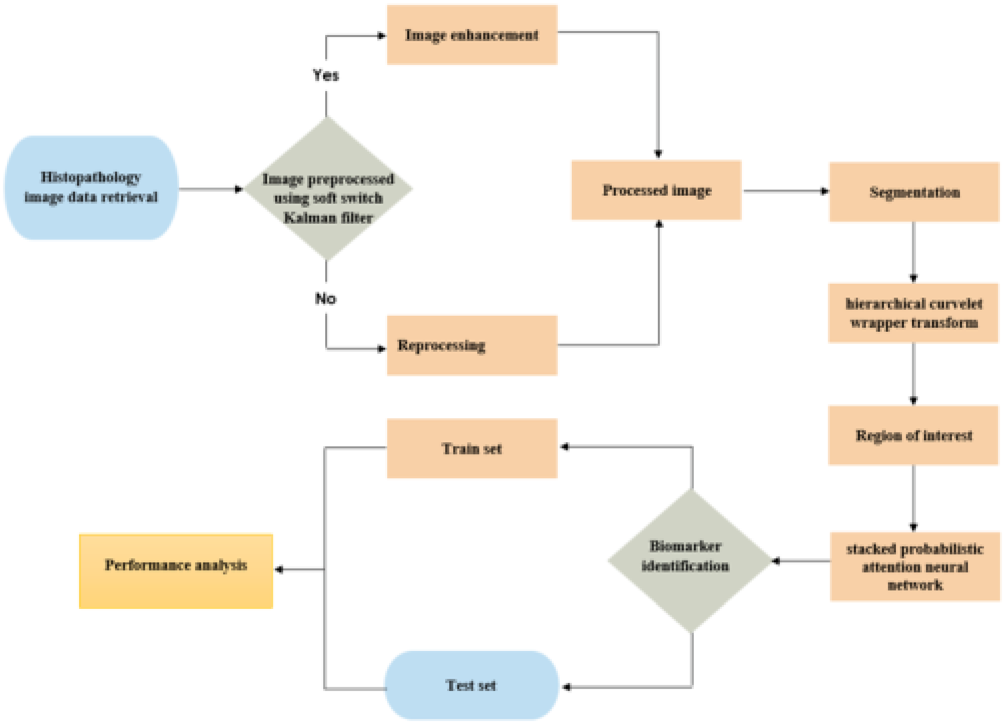

Histopathology image data retrieval
Image preprocessed using soft switch Kalman filter
Yes
No
Image enhancement
Reprocessing
Processed image
Segmentation
hierarchical curvelet wrapper transform
Region of interest
stacked probabilistic attention neural network
Biomarker identification
Train set
Test set
Performance analysis

Fig. 4.1: Schematic representation of the suggested methodology

4. Proposed methodology. Major advances in image processing and learning methodologies have not yet removed significant barriers to the development of quantitative imaging biomarkers for use in medical decision making.The recommended strategy for predicting the NHL-related biomarkers is shown in Figure 4.1.

4.1. Dataset. The dataset was extracted from the Kaggle https://www.kaggle.com/datasets/sparxi/ihc-images. It contains 28.8k IHC images for biomarker prediction.

4.2. Preprocessing. Images that have degraded and are noisy are filtered before being restored using various SSEF methods. The equivalent mathematical expression would be:

$$h(O,z) = f(O,z) * u(0,z) + n(O,z) \tag{4.1}$$

$$h(O,z) = R[h(O,z)] \tag{4.2}$$

The variables in the equation are defined as follows: f(O,z) represents the original input picture, u(O,z) represents the degradation function, "*" denotes the error function, n(O,z) represents the noise (usually Gaussian noise), g(O,z) represents the deteriorated output image, and h(O,z) represents the final degraded output image after the application of procedure R. By using noise reduction filters that use nonlinear spatial domains, like the one seen in this example, it becomes feasible to reconstruct denoised images from noisy source images. Here are few methods to improve the quality of your photographs: The first step in the noise-reduction filter involves creating a mask matrix of dimensions nm. The mask pixel value and mask pixel size are used in the mask matrix to calculate the new pixel value for the degraded image. The filter assigns the value of each pixel to be equal to the element in the middle of the mask's matrix. This method has the capability to eliminate irregularities without compromising the quality of the image. The proposed filter use the average and standard deviation of the pixel values in the mask matrix.

$$\mu = \frac{1}{OM} \sum_{n,m \in \eta} a(0,m) \tag{4.3}$$

$$\sigma^2 = \frac{1}{oM} \sum_{0,m \in o} a^2(o,m) - \mu^2 \tag{4.4}$$

The mask's neighbourhood area has a size of nm., σ^2 is the variance of the Gaussian noise in the image, and $a(n,m)$ is the representation of each pixel in the mask. Next, the SSEF filter is created for the updated pixels using the expected values, which are represented as $b_w(o,m)$.

$$c_w(o,m) = \mu + \frac{\sigma^2 - v^2}{\sigma^2}.(a(0,m) - \mu) \tag{4.5}$$

wherev^2 is the mask matrix's noise variance setting when using the SSEF filter.
Now, the imputed pixel values are given by

$$c_i^{imp} = \sum_{j=1}^{k} w_j z_j, \ \ i = 1,..,m \tag{4.6}$$

After error removal and imputing the pixel values the error free images are obtained.

4.3. Segmentation. The HSPT segmentation algorithm may be fed the processed picture. We'll refer to the areas in S_i that contain the first seeds, or $B_1, B_2, , B_i$. $(\bar{O}, \bar{D}_b, \bar{D}_r)$ to show how the sum of all S_i seed pixels breaks down into O, D_b, andD_r. In this section, we outline our suggested method of segmentation.

(1) Choose your seeds automatically.
(2) Give each seed area a label.

The seed pixel, first, has to share a lot of characteristics with its surrounding pixels. Second, in order to construct the predicted area, at least one seed must be created. Third, it's important to keep seeds for various locations apart.

The following formula is used to calculate the degree of similarity between a given pixel and its neighbours. The dispersion measures of the Y, C_b, and C_r Using the, the components of a 3×3

$$\sigma_Y = \sqrt{\frac{1}{9}\sum_{i=1}^{9}(O_i - \bar{O})^2}, \tag{4.7}$$

Where O can be Y, D_b, or D_r, the mean value $\bar{Y} = \frac{1}{9}\sum_{i=1}^{9} x_i$. Standard deviation, on the whole, is

$$\sigma = \sigma_K + \sigma_{D_b} + \sigma_{D_\tau} \tag{4.8}$$

To get the standard deviation back inside the range $[0,1]$, we,

$$\sigma_M = \frac{\sigma}{\sigma_{\max}}, \tag{4.9}$$

where $\sigma_{\max}$ is the image's greatest standard deviation. We may define a pixel's resemblance to its neighbours as

$$H = 1 - \sigma_M \tag{4.10}$$

The first requirement for the potential seed pixel is derived from the degree of similarity as follows.

The threshold similarity of a seed pixel candidate must be greater than 1.

The second step is to determine the YD_bD_r distances (relative Euclidean distances) between a pixel and its immediate neighbours.

$$d_i = \frac{\sqrt{(O - O_i)^2 + (D_b - D_{bi})^2 + (D_r - D_{r_1})^2}}{\sqrt{O^2 + D_b^2 + D_r^2}} \tag{4.11}$$

where $i = 1, 2, ...8$.

We determine the greatest possible separation between each pixel and its neighbours as,

$$d_{\max} = \max_{i=1}^{8}(d_i) \tag{4.12}$$

Create a list T of all the areas that are close by, then sort them by decreasing distance.

Remove the first point (p), even if T is not empty, and check to see whether any of its four neighbor's are empty. If all of p's labelled neighbor's have the same label, then p ought to get that label as well. If p's labelled neighbor's have different labels, p should be put in the area that is closest to it. The region's mean is then adjusted, and T is then expanded to include p's unclassified neighbor's in decreasing order of distance.

Using this method, we may get the fractional Euclidean distance, di, between pixel i and its neighbours.

$$d_i = \frac{\sqrt{(Z_i - \bar{Z})^2} + (D_{b_i} - D)^2 + (D_D - \bar{D}_r)^2}{\sqrt{Z_i + D_{b_i}^2 + D_{r_i}^2}} \tag{4.13}$$

where $(\bar{O}, \bar{D}, \bar{D}_r)$ are the medians of the distributions of Y, D_b, and D_r in the region. Pixel with the shortest distance value, p, is selected as the best one. If several neighbouring pixels have the same minimum value, we choose the one that best characterises the bigger of the two adjacent areas.

The red pixels indicate seeds, green pixels represent pixels in a sorted list T, white pixels represent the pixels with the shortest distance to the seed areas, white pixels are linked to the surrounding red region, and black pixels are added to which causes a recalculation of the mean of the new region and the distances between the new region and its neighbours. Once there is nowhere left where the distance is less than the criteria, we stop. What the distance between two points is in Euclidean space.

When discussing the colour differences between these regions, we use the labels R_i and R_j.

$$d_i = \frac{\sqrt{(\bar{O}_i - \bar{O}_j)^2 + (\bar{C}_D - \bar{C}_D)^2 + (\bar{D}_{r_i} - \bar{D}_{r_j})^2}}{\sqrt{O^2 + D_b^2 + D_r^2}} \tag{4.14}$$

After repeating the process the ROI can be separated.

4.4. Biomarker Identification. SPANN uses a direct influence on network structure data to identify unlabeled nodes by transmitting their labels across transfer and sink nodes. Equation 4.15 defines undirected graphs.

$$\zeta = (G, \epsilon) \tag{4.15}$$

where $G = G_1, G_2, , G_n$ represent nodes, $\epsilon = \epsilon_1, \epsilon_2, , \epsilon_n$ represent edges. Matrix adjacency ζ, A' may be calculated to determine whether two nodes are related

$$B_{ij}^{'} = \begin{cases} \alpha, & G_i \neq G_j \\ 1, & G_i = G_j \\ e^{K||G_i - G_j||}, & otherwise. \end{cases} \tag{4.16}$$

where, α_{mhsa_ij} value of 0.2 in the studies, demonstrating multi-head self-guided attention determines neighbour node weights.

We provide multi-tiered SPANN topologies. This data allows real-time adjacency matrix adjustments.

$$B^{'(r)} \leftarrow B(B^{'(r-1)} + \alpha V^{(r-1)} h^{(r)T}) B^T + \beta l \tag{4.17}$$

where $I^{(r-1)}$ represents biomarker-associated aspects of the $(r-1)"th"$ layer's output; indicates the coefficient of correlation. $h^{(r)}$ indicates represents biomarker-associated aspects of the $(r-1)"th"$ layer's output that is the coefficient of correlation

$$l^{(r)} = \delta(BX^r) \tag{4.18}$$

where δ signifies the softplus $(.)$ activation function is engaged, and $\bar{B}$ may take the values specified by Equation (4.19).

$$\begin{cases} 1+E^{-\frac{1}{2}}B^{'}E^{-\frac{1}{2}}\tilde{\rightarrow}E^{\frac{1}{2}}\tilde{A}E^{-\frac{1}{2}} \\ \tilde{E_{ij}}=\sum_{j}\tilde{A_{ij}} \end{cases} \tag{4.19}$$

I is the identity matrix.

We use the stack attention module to reduce superfluous subspace pixel blocks and create the same pixel block with the completely connect pixel cut to carry different numbers of subpixel block nodes depending on size. To divide subpixel blocks optimally, compute the global and local property information gain. Equation 4.20 shows subpixel block formation.

$$\beta_{t,x}=\lambda\beta_{t,x}^{global}+\gamma\beta_{t,x}^{local},\ \ \gamma=\frac{1}{2}-\lambda \tag{4.20}$$

where γ,λ indicates weighing factor $\beta_{t,x}^{global}$ and $\beta_{t,x}^{local}$ signals global or region attention $\beta_{t,x}^{global}$ and $\beta_{t,x}^{local}$ as:

$$\begin{cases} \beta_{t,x}^{global}=\frac{\exp(score(V_t,\bar{V_x}))}{\sum_{Y=1}^{T_Y}\exp(score(V_t,\bar{V_x}'))} \\ \beta_{t,x}^{local}=\frac{\exp(score(h_t,\bar{h_x}))}{\sum_{T_x\delta(v_p^T tanV(w_pv_t))-E}^{T_x\delta(v_p^T tanV(w_pv_t))+E}\exp(score(V_t,V_x'))}exp\left(-\frac{Y-T_Y.\delta(v_p^T tanV(W_pV_t))}{8E^2}\right) \end{cases} \tag{4.21}$$

where δ activates a function.

M distinct attention-directed adjacency matrices need M tightly connected layers. We alter each layer's calculation as stated below (for the $l"th"$ matrix $\bar{A}^t$) in equation 22 .

$$V_{ti}^{l}=\delta(\sum_{j=1}^{n}\bar{B}_{ij}^{t}W_t^{(l)}Z_j^{(l)}+b_t^{(l)}) \tag{4.22}$$

Focused adjacencies $\bar{A}^t.Z_j^{(l)}$, where t=1,...,M, and t selects the bias term and weight matrix related to $\bar{A}^t$.

$$Z_j^{(l)}=[X_j,V_j^{(l)},...,V_j^{(l)}] \tag{4.23}$$

L is the number of closely coupled layer sublayers. The stack's primary purpose is to split the super pixel block of flawlessly connected pixels into a smaller subspace to create an adjacency matrix. Since a connectionless edge's weight is set to 0, pruning is necessary. Cutting the completely connected super pixel block might destroy part-relevant information. Thus, we designed a self-attention guidance module to redistribute edge weight to the trimmed subspace pixel block, stressing graph node relationships and interactions, establishing a more dependable multi-scale graph structure, and addressing the problem.

The self-guided attention module transforms the multi-scale subspace pixel block into a totally linked graph using multi-head self-attention. While attention guidance builds an adjacency matrix A', edge weights are enhanced. Each A' represents a totally connected graph, and entry Aij' denotes the degree of connection between nodes i and j. Attention to build node relations allows the self-attention machine to record interactions between any two places in a single sequence. Equation 4.4 calculates A'.

$$\alpha_{SPANNs_ij}=\frac{exp(LeakyReLu(\bar{a}^T[w\bar{v_i}||w\bar{v_j}]))}{\sum_{k\in N_i}exp(LeakyReLu(\bar{a}^T[w\bar{v_i}||w\bar{v_k}]))} \tag{4.24}$$

where T represents the matrix transpose and w the node weights. Node i's neighbours, denoted by N_i, are i, LeakyReLu(.) indicates activate function.

For example, we may join the results of the recommended network, which has a fully-connected layer for the masked function, as.

$$FCN_{g_{st}}=SoftMax([HCN_{g_s0},....HCN_{g_st}]W_{fcn}) \tag{4.25}$$

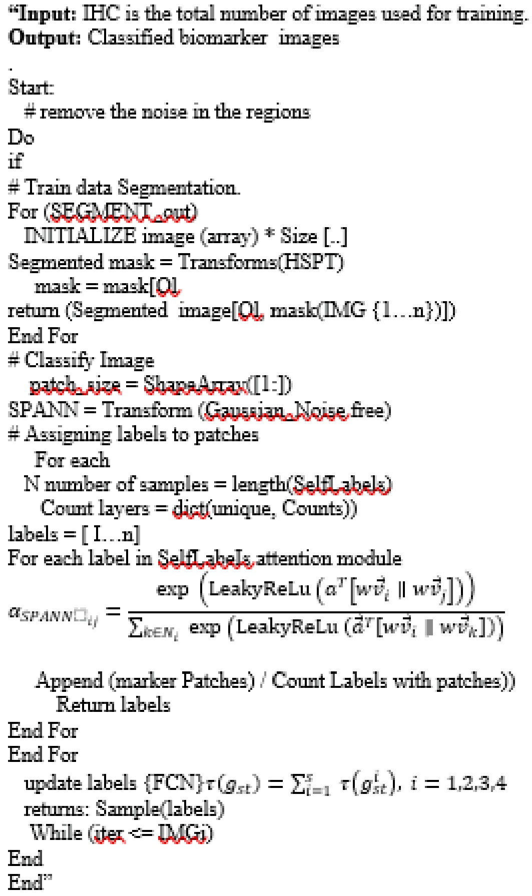
where $st = 0, 1, 2, 3, W_f cn$ represents layer weights when all nodes are linked.

$$\zeta(V_{st}) = \sum_{i=1}^{s} \zeta(V_{st}^{i}), \quad i = 1, 2, 3, 4 \tag{4.26}$$

where $\zeta(V_st)$ i denotes the cross-entropy error loss, which is a measure of how far a network's predictions deviate from the labels used to build the training set.

Algorithm 9 illustrates the implementation procedure of the SPANN.

Algorithm 9 SPANN



```
"Input: IHC is the total number of images used for training.
Output: Classified biomarker  images
.
Start:
  # remove the noise in the regions
Do
if
# Train data Segmentation.
For (SEGMENT_out)
  INITIALIZE image (array) * Size [..]
Segmented mask = Transforms(HSPT)
    mask = mask[O],
return (Segmented  image[O], mask(IMG {1…n})])
End For
# Classify Image
   patch_size = ShapeArray([1:])
SPANN = Transform (Gaussian_Noise free)
# Assigning labels to patches
    For each
  N number of samples = length(SelfLabels)
     Count layers = dict(unique, Counts))
labels = [ I…n]
For each label in SelfLabels attention module
$\alpha_{SPANN\square_{ij}} = \frac{\exp\left(\text{LeakyReLu}\left(a^{T}[w\vec{v}_i \parallel w\vec{v}_j]\right)\right)}{\sum_{k\in N_i} \exp\left(\text{LeakyReLu}\left(\vec{a}^{T}[w\vec{v}_i \parallel w\vec{v}_k]\right)\right)}$

    Append (marker Patches) / Count Labels with patches))
      Return labels
End For
End For
  update labels {FCN}$\tau(g_{st}) = \sum_{i=1}^{s} \tau(g_{st}^{i})$, $i = 1,2,3,4$
  returns: Sample(labels)
   While (iter <= IMGi)
End
End"
```

5. Performance analysis. Here, we provide empirical data that substantiates the efficacy of the suggested analytical methodology. In general, the tests were carried out in a Python environment. The parameters of the proposed solution for biomarker prediction are computed, and the system's efficiency is compared to that of current techniques.

Figure 5.1 depicts a visual representation of the sample input acquired from the Kaggle database.

The objective of preprocessing is to optimise the efficiency of the classifier by determining the most valuable set of features. In this scenario, the Gaussian error in the picture may be repaired, as seen in Figure 5.2.

The objective of HSPT picture segmentation is to assign a categorical label to every individual pixel. We are using pixel-level predictions to identify the Region of Interest (ROI) within the selected parts of the image. Figure 5.3 displays the segmented output.

SPANN was used to examine the IHC protein markers included in the datasets under scrutiny. Every point is an immunohistochemical (IHC) picture of a marker found in the NHL. Figure 5.4 displays illustrative

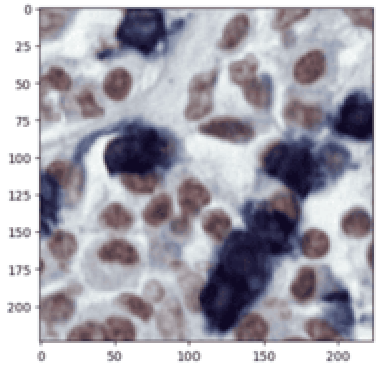


Fig. 5.1: Sample input

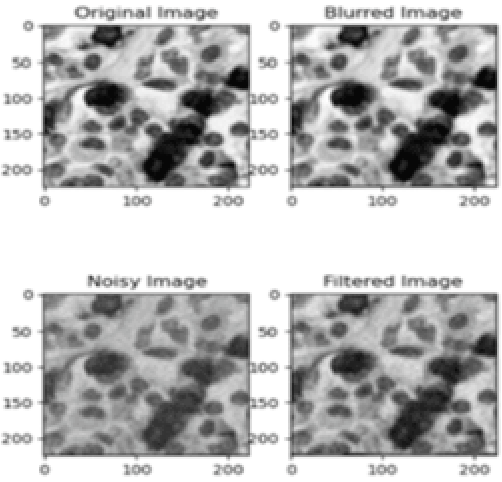


Fig. 5.2: Processed output

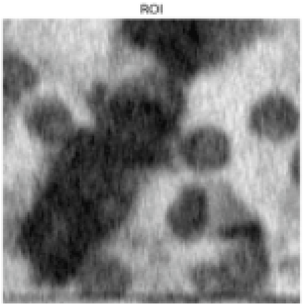


Fig. 5.3: Segmented output

examples of each marker picked at random. Figure 5.4 shows that the proposed technique can effectively separate and categorize the diverse group of testing sets spanning three distinct IHC markers.

Training accuracy and efficiency depend on epochs. Too few epochs may not provide the model enough

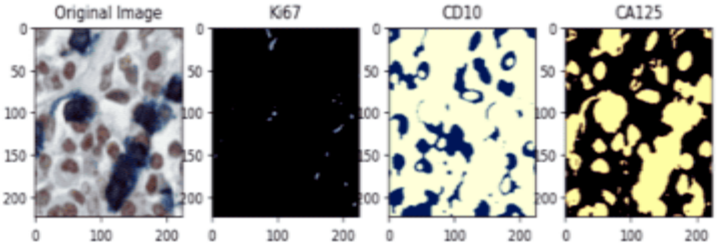


Fig. 5.4: Output classification

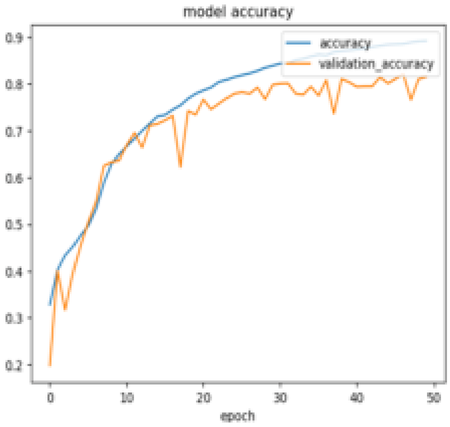


Fig. 5.5: Epoch vs. Accuracy

time to comprehend the data structure. Increase epoch size to 0.95 percent to increase accuracy. Similar binary classification solutions are assessed using confusion matrices. The confusion matrix evaluates categorization solutions by comparing predictions to reality. Displays false negatives, accurate forecasts, and incorrect predictions. Confusion matrix-based classifier assessment metrics may be constructed from this data. SPANN and learning-based models are evaluated using accuracy, recall, precision, F-measure, and AUC.

Accuracy. This heuristic performance metric predicts accuracy. Equation (5.1) calculates score by dividing total occurrences by accurate guesses.

$$Recall = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Recall. It is sometimes referred to as sensitivity. Equation (5.2), which, when solved, gives the percentage of outcomes that were properly predicted when the result was positive, may be used to calculate this metric.

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

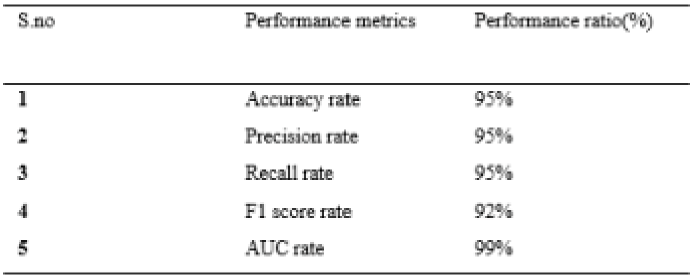
Precision. It is the ratio of accurately anticipated positive occurrences to the total number forecasted. Its formula can solve (29).

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

F-measure. When class sizes are uneven, it is a common performance measure. This measure averages accuracy and recall scores, as stated in Equation (5.4).

$$F - Measure = 2 \times \frac{Prec \times Rec}{Prec + Rec} \tag{5.4}$$

Table 5.1: Performance analysis



S.no	Performance metrics	Performance ratio(%)
1	Accuracy rate	95%
2	Precision rate	95%
3	Recall rate	95%
4	F1 score rate	92%
5	AUC rate	99%

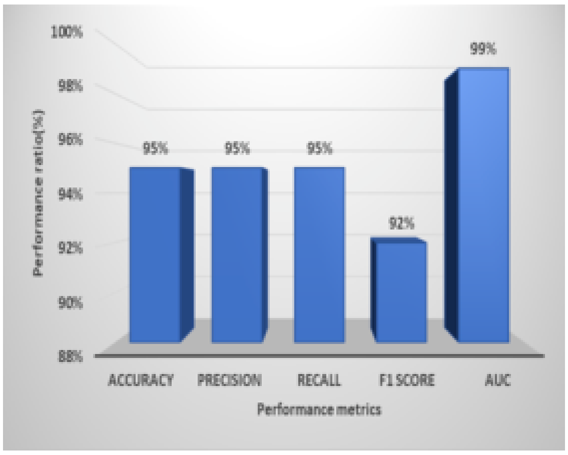


Fig. 5.6: Performance analysis of the suggested methodology

Area under the curve. Measures model categorisation accuracy. Equation (20) estimates the area under the receiver operating characteristic (ROC) curve for test evaluation.

$$AUC = \int TruepositivityR, d(FalPR) \tag{5.5}$$

TPR and FPR. Integrating the TPR with regard to the FPR yields the AUC score from the area under the ROC curve, which demonstrates the connection between these ratios.

There are multiple matches for performance evaluation methodology evaluation, including performance evaluation methods and performance evaluation process . Here we are evaluating our suggested mechanism with accuracy, precision, recall and F score. Table 5.1 and figure 5.6 show the methodology's performance. Comparing the proposed technique to known mechanisms helps assess its efficacy[22,11,21].

Divide the total of all true positives and negatives by the sum to get a classifier's accuracy. The suggested approach is 95 percentage more accurate than traditional practises (see Figure 5.7).

To obtain the precision for a given class, we divide the number of true positives by the classifier bias towards this class (number of times that the classifier has predicted the class). Figure 5.8 shows that HSPT and SPANN (95 percentage) outperform other biomarker prediction techniques.

Based on Figure 5.9, the suggested HSPT and SPANN technique has a recall of up to 95 percentage, which is far higher than existing approaches.

Commonly used as an evaluation metric in binary and multi-class classification , the F1 score integrates precision and recall into a single metric to gain a better understanding of model performance The proposed

Table 5.2: Comparative performance analysis

Methodology	Accuracy	Precision	AUC	F1	Recall
LR [22]	0.869	0.871	0.887	0.803	0.762
Adaboost [22]	0.808	0.714	0.806	0.722	0.747
Decision Tree [22]	0.812	0.790	0.806	0.708	0.665
Boost [22]	0.842	0.822	0.875	0.759	0.720
SVM [22]	0.849	0.932	0.890	0.747	0.63
Resnet 18 [11]	0.93	0.95	-	0.937	0.937
Proposed	**0.95**	**0.95**	**0.99**	**0.95**	**0.95**

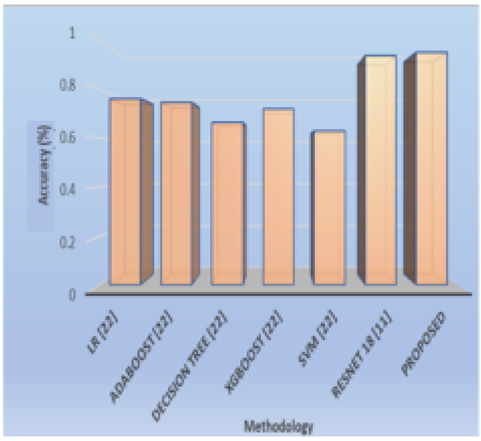


Fig. 5.7: Accuracy percentile analysis

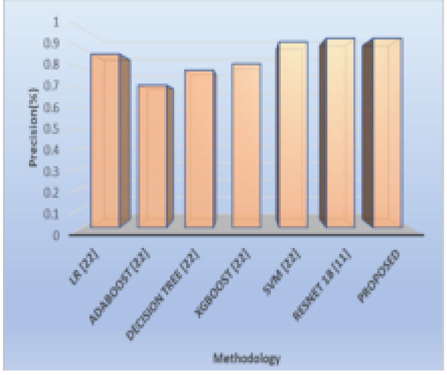


Fig. 5.8: Precision percentile analysis

HSPT and SPANN approach has a high rate of F1 score (95 percentage) compared to the current mechanisms, as can be shown in Figure 5.10.

Figure 5.11 depicts the two-dimensional ROC curve. The x-axis indicates positive rates, the y-axis shows true positive rates, and the threshold ranges from 0 to 1 (higher right to lower left). All threshold classification

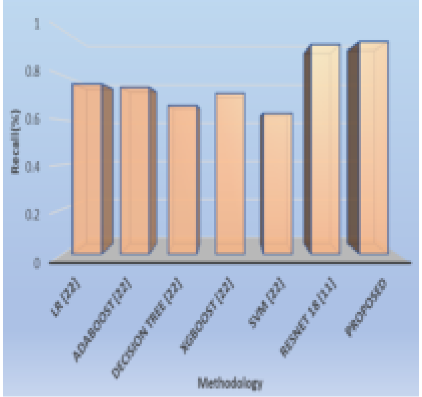


Fig. 5.9: Recall percentile analysis

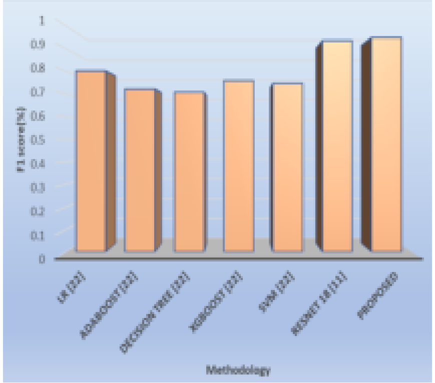


Fig. 5.10: F1 score percentile analysis

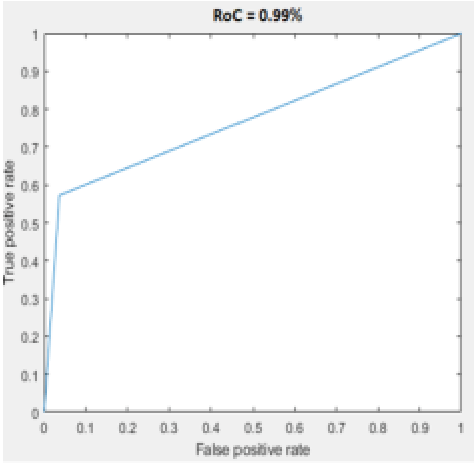
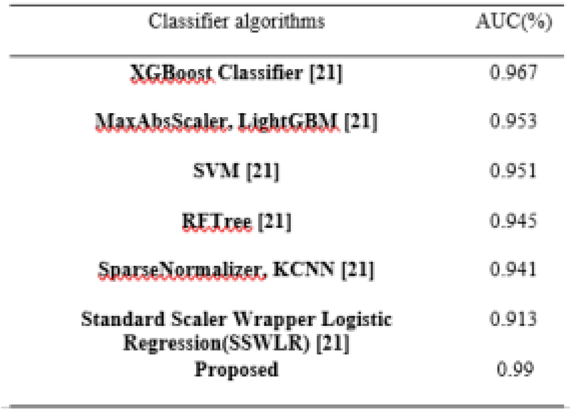


Fig. 5.11: ROC analysis

Table 5.3: AUC analysis



Classifier algorithms	AUC(%)
XGBoost Classifier [21]	0.967
MaxAbsScaler, LightGBM [21]	0.953
SVM [21]	0.951
RFTree [21]	0.945
SparseNormalizer, KCNN [21]	0.941
Standard Scaler Wrapper Logistic Regression(SSWLR) [21]	0.913
Proposed	0.99

results are graphed. An AUC of 99 percentage implies that the classifier is completely accurate.

Table 5.3 of the AUC performance measure comparison demonstrates that DL distinguishes biomarkers well. Our HSPT and SPANN models outperform gold standard methods. The table proves the suggested model is better than its competitors. The suggested model outperforms Tables 5.2 and 5.3. Compared to previous biomarker prediction mechanisms, the recommended technique yields satisfactory results".

6. Conclusion. Slides stained for Ki-67, CD 10, and CA125 were analysed to see whether they might be used to predict outcomes for NHL patients. To anticipate biomarker expression from H& E stained pictures without the need for IHC labelling, we developed an HSPT and SPANN model. Our findings demonstrate the close relationship between morphological and molecular data by demonstrating that histological pictures of tissue and cell morphologies have underlying molecular origins. Once this connection has been discovered, the abundance of a target protein may be predicted among the samples using a deep learning-based technique. The proposed strategy here significantly beat the state-of-the-art biomarker prediction mechanisms, by as much as 95 percentage. The following are where our future efforts will be concentrated. We need to enlarge our sample size to get more accurate results. By training the model on the new data, its resilience and generalization abilities will increase. To further generalize our findings, we recommend further trials on samples including a variety of tissues and stains; also, there is a suggestion for optimizing the model. Semi-supervised learning, for instance, may be used to reduce the burden of annotation. While our work demonstrated the capability of tumour histology to forecast pCR using DL methodologies and introduced a unique biomarker that serves as a more efficacious predictor than sTILs or subtype, it remains subject to certain limitations. This study used a restricted number of patients retrospectively for training and validation; hence, future research should aim for prospective multicenter investigations.

Acknowledgment. I would like to express profound gratitude to my guide for this invaluable support, encouragement, supervision and useful suggestions throughout this study. Her moral support and continuous guidance enabled me to complete my work successfully.

Author's Contributions. *Sivaranjini N.:* Designed, analysis and acquisition of data. *Gomathi M.:* Reviewed and Organized the study.

REFERENCES

[1] J. Carreras, Y. Y. Kikuti, M. Miyaoka, S. Hiraiwa, S. Tomita, H. Ikoma, et al., "A combination of multilayer perceptron, radial basis function artificial neural networks and machine learning image segmentation for the dimension reduction and the prognosis assessment of diffuse large B-cell lymphoma," AI, vol. 2, pp. 106-134, 2021.

[2] D. Vrabac, A. Smit, R. Rojansky, Y. Natkunam, R. H. Advani, A. Y. Ng, et al., "DLBCL-Morph: Morphological features computed using deep learning for an annotated digital DLBCL image set," Scientific Data, vol. 8, p. 135, 2021.

[3] I. S. Lossos, ”Diffuse large B cell lymphoma: from gene expression profiling to prediction of outcome,” Biology of blood and marrow transplantation, vol. 14, pp. 108-111, 2008.

[4] J. Perdiz Arrais and J. L. Oliveira, ”RecRWR: a recursive random walk method for improved identification of diseases,” BioMed Research International, vol. 2015, 2015.

[5] F. Aziz, A. Acharjee, J. A. Williams, D. Russ, L. Bravo-Merodio, and G. V. Gkoutos, ”Biomarker prioritisation and power estimation using ensemble gene regulatory network inference,” International Journal of Molecular Sciences, vol. 21, p. 7886, 2020.

[6] K. Vaishali, S. Shambharkar, R. K. Somkunwar, and R. R. Kolte, ”Augmented Ensemble Learning Model for Biomarkers Prioritization to Enhance Disease Identification Efficiency,” in 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 2023, pp. 1-7.

[7] R. Hajjo, D. A. Sabbah, S. K. Bardaweel, and A. Tropsha, ”Identification of tumor-specific MRI biomarkers using machine learning (ML),” Diagnostics, vol. 11, p. 742, 2021.

[8] Musthafa, M. M., TR, M., V, V. K., & Guluwadi, S. (2024). Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification. BMC Medical Imaging, 24(1), 201

[9] Y. Chen, H. Yang, Z. Cheng, L. Chen, S. Peng, J. Wang, et al., ”A whole-slide image (WSI)-based immunohistochemical feature prediction system improves the subtyping of lung cancer,” Lung Cancer, vol. 165, pp. 18-27, 2022.

[10] T. E. Tavolara, M. K. K. Niazi, D. Jaye, C. Flowers, L. Cooper, and M. N. Gurcan, ”Deep learning to predict the proportion of positive cells in CMYC-stained tissue microarrays of diffuse large B-cell lymphoma,” in Medical Imaging 2023: Digital and Computational Pathology, 2023, pp. 12-16.

[11] Y. Liu, X. Li, A. Zheng, X. Zhu, S. Liu, M. Hu, et al., ”Predict Ki-67 positive cells in H& E-stained images using deep learning independently from IHC-stained images,” Frontiers in Molecular Biosciences, vol. 7, p. 183, 2020.

[12] C. Syrykh, A. Abreu, N. Amara, A. Siegfried, V. Maisongrosse, F. X. Frenois, et al., ”Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning,” NPJ digital medicine, vol. 3, p. 63, 2020.

[13] Chakravarthy, S., Nagarajan, B., Kumar, V. V., Mahesh, T. R., Sivakami, R., & Annand, J. R. (2024). Breast tumor classification with enhanced transfer learning features and selection using chaotic map-based optimization. International Journal of Computational Intelligence Systems, 17(1), 18.

[14] P. Ghahremani, Y. Li, A. Kaufman, R. Vanguri, N. Greenwald, M. Angelo, et al., ”Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification,” Nature machine intelligence, vol. 4, pp. 401-412, 2022.

[15] V. Bobée, F. Drieux, V. Marchand, V. Sater, L. Veresezan, J.-M. Picquenot, et al., ”Combining gene expression profiling and machine learning to diagnose B-cell non-Hodgkin lymphoma,” Blood Cancer Journal, vol. 10, p. 59, 2020.

[16] C. Beaulac, J. S. Rosenthal, Q. Pei, D. Friedman, S. Wolden, and D. Hodgson, ”An evaluation of machine learning techniques to predict the outcome of children treated for Hodgkin-Lymphoma on the AHOD0031 trial,” Applied Artificial Intelligence, vol. 34, pp. 1100-1114, 2020.

[17] J. Zhang, W. Cui, X. Guo, B. Wang, and Z. Wang, ”Classification of digital pathological images of non-Hodgkin's lymphoma subtypes based on the fusion of transfer learning and principal component analysis,” Medical Physics, vol. 47, pp. 4241-4253, 2020.

[18] J. Carreras and R. Hamoudi, ”Artificial neural network analysis of gene expression data predicted non-hodgkin lymphoma subtypes with high accuracy,” Machine Learning and Knowledge Extraction, vol. 3, pp. 720-739, 2021.

[19] Ahmed, S. T., Sivakami, R., Mahesh, T. R., Khan, S. B., Mashat, A., & Almusharraf, A. (2024). PrEGAN: Privacy Enhanced Clinical EMR Generation: Leveraging GAN Model for Customer De-Identification. IEEE Transactions on Consumer Electronics..

[20] J. J. Eertink, G. J. Zwezerijnen, S. E. Wiegers, S. Pieplenbosch, M. E. Chamuleau, P. J. Lugtenburg, et al., ”Baseline radiomics features and MYC rearrangement status predict progression in aggressive B-cell lymphoma,” Blood Advances, vol. 7, pp. 214-223, 2023.

[21] García, R., Hussain, A., Chen, W., Wilson, K., & Koduru, P. (2022). An artificial intelligence system applied to recurrent cytogenetic aberrations and genetic progression scores predicts MYC rearrangements in large B-cell lymphoma. EJHaem, 3(3), 707-721.

[22] Hao, P., Deng, B. Y., Huang, C. T., Xu, J., Zhou, F., Liu, Z. X., ... & Xu, Y. K. (2022). Predicting anaplastic lymphoma kinase rearrangement status in patients with non-small cell lung cancer using a machine learning algorithm that combines clinical features and CT images. Frontiers in Oncology, 12, 5627.

[23] Bharanidharan, N., Chakravarthy, S. S., Venkatesan, V. K., Abbas, M., Mahesh, T. R., Mohan, E., & Venkatesan, K. (2024). Local entropy based remora optimization and sparse autoencoders for cancer diagnosis through microarray gene expression analysis. IEEE Access..

[24] Thakur, A., Gupta, M., Sinha, D. K., Mishra, K. K., Venkatesan, V. K., & Guluwadi, S. (2024). Transformative breast Cancer diagnosis using CNNs with optimized ReduceLROnPlateau and Early stopping Enhancements. International Journal of Computational Intelligence Systems, 17(1), 14..

[25] Mahesh, T. R., Vinoth Kumar, V., Vivek, V., Karthick Raghunath, K. M., & Sindhu Madhuri, G. (2024). Early predictive model for breast cancer classification using blended ensemble learning. International Journal of System Assurance Engineering and Management, 15(1), 188-197..

[26] Z. L. . Huan Wang , ”Using Machine Learning to Expand the Ann Arbor Staging System for Hodgkin and Non-Hodgkin Lymphoma,” BioMedInformatics, p. 12, 2023.

[27] Bruce D. Cheson et al, ”Recommendations for Initial Evaluation, Staging, and Response Assessment of Hodgkin and Non-Hodgkin Lymphoma: The Lugano Classification,” Journal Of Clinical Oncology, vol. 32, p. 10, 2024.

[28] Mahesh, T. R., Geman, O., Margala, M., & Guduri, M. (2023). The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. Healthcare Analytics, 4, 100247..

[29] Andy N.D. Nguyen, Kareem A. Allam, ”Deep Learning for Digital Image Analysis with Whole Slide Imaging for Lymphoma Diagnosis: Challenges and Promises,” 21st Century Pathology, p. 10, 2022.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jun 16, 2024
Accepted: Oct 2, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1165–1179, DOI 10.12694/scpe.v26i3.4331

ENERGY EFFICIENCY TASK RE-SCHEDULING IN VIRTUALIZED CLOUD COMPUTING

P.HANUMANTHA RAO[*], P.S. RAJAKUMAR[†], AND S. GEETHA[‡]

Abstract. Reduced energy consumption is an important goal for virtualized cloud computing systems since it has the potential to improve system efficiency, save operating costs, and lessen environmental impact. These objectives can be achieved by using an energy-efficient approach to job scheduling. The huge challenge lies in coordinating user demands with available cloud resources in a way that maximizes performance while reducing energy usage, all within the time frame that the user specifies. This article suggests a novel method called Energy Efficient Task Re-scheduling (EETRS) for a heterogeneous virtualized cloud environment as a solution to the problem of energy usage. The first step of the suggested approach assigns jobs strictly according to due dates, ignoring energy consumption. Task reassignment scheduling determines the optimal execution location within the deadline constraints while minimizing energy consumption in the second stage of the proposed method, which speeds up execution and meets deadlines. According to the simulation results, the suggested technique helps to significantly reduce energy use and boost performance by 5% while satisfying deadline constraints, in comparison to the current energy-efficient scheduling methods of EPETS, AMTS, and EPAGA. The proposed method outperforms the existing one with less than 1% total execution time, a reduction of 14% in total execution cost, a 3% decrease in energy consumption, and a 3% reduction in average resource utilization.

Key words: Cloud Computing, Entergy Consumption, Task Re-scheduling, Energy minimization, Performance enhancement

1. Introduction.

Computing entered a new age with the advent of cloud computing, as a result of technological advancements that integrated storage, processing power, and networks. The term ”cloud computing” refers to a new face of shared computing that enables users to gain access to shared computer resources whenever they need them through an internet connection. Providers of cloud computing services that host several applications have several responsibilities, including adhering to service-level agreements (SLAs), ensuring reliable and secure data management, meeting task deadlines, and achieving low access latencies. Commercial goals of cloud providers may conflict with energy-efficient, cost-effective hardware designs and capacity planning strategies used by back-end data centers.

Data center energy management is intricate because it requires real-time evaluation of dynamic factors such as traffic conditions, inter-process communication, workload and resource allocation, and cooling plans. With the market for cloud pricing becoming more intensely competitive, cloud companies are under increasing pressure to find ways to power down their data centers' backend [1]. The typical data center workload is around 30% and does not require full computer resources to be used [2]. Consequently, it is possible to match the workload demands of the data center while saving energy by turning off certain idle equipment. But data replication, client SLAs, performance, and latency issues, and data center traffic patterns [3] must all be carefully considered when data center resource scheduling is to be performed.

Critical and energy-intensive, data centers deliver Internet-based services on a massive scale. To reduce excessive energy use in data centers, power utilization models are crucial for developing and improving energy-efficient operations. The rapid expansion of distributed cloud computing network services has led to an explosion of data sizes across various industries, encompassing signal processing, bio-informatics, IoT, and scientific computing. Cloud computing applications make use of the thousands of powerful servers hosted in cloud data centers to execute millions of jobs. Virtualization is one of the greatest advantages for consumers who can

[*] Dept. of CSE, Dr.M.G.R. Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu, India (hanumanthraovbit@gmail.com)

[†] Dept. of CSE ,Dr.M.G.R. Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu, India (Rajakumar.subramanian@drmgrdu.ac.in)

[‡] Dept. of CSE, Dr.M.G.R. Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu, India (somangeetha@drmgrdu.ac.in)

take advantage of the many services offered by the cloud through virtual machines (VMs). A lot of power is typically used by these virtual machines. The cost of electricity goes up and the environment suffers as a result of this kind of energy usage [4]. Despite technological efforts to reduce energy use, the data center is predicted to discharge 62 million tonnes of carbon dioxide (CO2) into the environment [5]. Low utilization of computing resources and inefficient job scheduling are primary causes of enormous energy consumption [6]. In addition, the duties included in each application are diverse and might range in size. So, it's critical to manage energy consumption in cloud data centers and make them more energy efficient when scheduling. Energy efficiency in data centers is becoming increasingly important and complex in recent years. To keep data accessible at all times, the data center's various components must work together to reduce energy consumption and downtime. All information technology (IT) infrastructures are based on the technical infrastructure, which includes power supply, technical cooling, and technical security.

Presently, the vast majority of studies just allocate tasks to Virtual Machines (VMs) without taking into account the fact that different jobs have variable resource requirements [7]. A tremendous increase in energy use and a loss of valuable resources ensued. Regardless, getting optimal energy consumption and scheduling all workloads to appropriate servers is a huge concern. Many researchers in the academic community have worked on these problems to lower the power consumption of cloud data centers. Specifically, the energy-efficient scheduling issue in a wide variety of devices was investigated in research [8], [9], [10], and [11]. Scheduling all tasks according to their quality of service criteria while minimizing power usage was the main objective. Evidence from studies [12], [13], and [14] shows that hybrid data centers can save energy by combining powerful and less powerful servers. At now, most data centers use a variety of servers with varying degrees of computational power and power consumption features [15]. This means that various virtual machines (VMs) will have different execution durations and energy usage when processing the same activity. If the work is scheduled to the correct server form, it can help improve system performance and save energy usage.

The processing time of the task is practically related to the energy consumption problem. There are several obstacles to overcome in order to determine how long each task should take to complete in response to the following study questions.

1. Assigning tasks to the correct machine at the right time is difficult to reduce energy usage. Energy usage and resource waste are both exacerbated by the host's tough assignment of tasks.
2. Assigning all tasks to the same virtual machine, or one that is slower, would cause processing times to increase. Consequently, the scheduler will make the execution time increase, which will disrupt the deadlines of jobs. If all tasks are sent to a faster virtual machine, processing times will be reduced, but more energy will be consumed.
3. Earliest finish times (EFT), Processing times, latest completion times (LFT), and a deadline need to be defined in order to allocate jobs to virtual machines. There would be different slack times due to different allocations. Assigning work to slower machines reduces slack time and energy usage, but there is an enormous number of combinations of possible task allocation possibilities. Thus, evaluating all the combinations is extremely difficult, and perhaps impossible.

This research took into account several different tasks that each had their quality of service criteria, such as deadline, workload size, and execution priority. Using a diverse set of resources to choose a virtual machine that both satisfies a task's quality of service needs and uses the least amount of energy is thus becoming an increasingly difficult scheduling challenge. Given the aforementioned rationale, this article proposes a method, Energy Efficient Task Re-scheduling (EETRS), for scheduling tasks in a heterogeneous virtualized cloud that minimizes both energy consumption and enhances performance. The goal is to meet task objectives without sacrificing service quality while minimizing energy consumption during scheduling. The paper's outline is as follows: A brief synopsis of relevant studies is given in the second section. Section 3 describes the algorithm and workings of the proposed approach EETRS. Section 4 presents the suggested method's experimental settings, results, and a thorough performance analysis. Section 5 outlines conclusions and future steps.

2. Literature Review. Recent years have seen virtualization technologies rise to the forefront of computer system architecture once again.

Secure computing, transparent migration, and consolidation of servers are just a few examples of the new capabilities that may be added to a system through virtual machines (VMs), which also allow developers to keep existing operating systems and applications compatible. Modern virtualized environments need all virtual machines (VMs) to share the same core to follow the hypervisor-controlled power management strategy. Various constraints apply to certain configurations. There is little opportunity for users to personalize the power management parameters for individual virtual machines. As a second point, it often impacts the energy efficiency of some or all virtual machines, especially when those VMs have different energy regulating plans that they require.

To address the challenges mentioned above, Li et al. [16] proposed a technique of power control that is specific to each virtual machine (VM). This method would allow each VM's guest operating system to use its preferred energy management strategy while simultaneously preventing similar VMs from competing with one another. Virtual performance (VIP) improves the timing of CPU-intensive applications by 32% and reduces power consumption by 27% in contrast to the Xen hypervisor's default on-demand governor, respectively, all without violating the service level agreement (SLA) of latency-sensitive implementations. The individual strategy of energy management is not possible in practice. In addition to optimizing energy efficiency by analysing the VM scheduling mechanism along with the virtualization paradigm of Input and output, Lee et al. [17] introduced a new offset mechanism to conquer fast input and output performance while power-fairness credit sequencing strategy. Also, virtual machine resource calibration was presented by Sheikh et al. [18]. They developed a method that uses power monitoring services and controlled feedback architecture to lower virtual servers' energy consumption. The above methods were inefficient both in minimizing the energy and enhancing the performance successfully.

Architectures typically provide several methods and processes for how to distribute and organize work across many resources using methods, including virtual machine placement, migration, consolidation, scheduling that minimizes energy use, and virtualization. A virtual machine placement problem was addressed by the authors in [19] through the use of an online meta-heuristic method that was dependent on the Ant Colony System. An objective function is used by the algorithm to find an approximation of the ideal solution. They were able to achieve better power usage without sacrificing the performance needed for a cloud data center. This method was less efficient, compared to other existing solutions. In their study, Arianyan et al. [20] implemented a method that helps improve resource utilization and performance while decreasing energy usage in cloud data centers. Energy usage, live migration, and SLA violations formed the basis of their new and effective resource management system. They broke the whole issue down into two smaller ones. After identifying the overworked server, it uses a multi-criteria selection decision approach to choose which virtual machine (VM) to move. However, this method introduces a lot of time complexity. On the other hand, a great deal of work has been done to determine how energy consumption relates to performance by the authors of [21]. In an effort to strike a better balance between power consumption and performance, they rethought the VM integration metric with energy efficiency. They established an SLA conflict algorithm inside this framework that takes into account minimal power consumption, maximum utilization, and SLA conflict as ways to identify when a server is overloaded. This effort aims to minimize energy usage without sacrificing service quality. This method incurs in overhead of large complex computations.

An analysis of the energy utilization of cloud computing was given in [22]. Public and private clouds were also taken into account in the study, along with the energy requirements of data computing, storage, and switching. They proved that a large amount of cloud computing's energy demand could be attributable to power consumption during transit and switching. Cloud computing (CC) is seen by their proposed approach as a counterpart to a traditional supply chain and logistics problem that accounts for the cost or power consumption of processing, storing, and transferring physical items. In addition, Yang Qin et al. [23] use an ILP model based on task profile information to optimize energy consumption for real-time activities using the intra-task DVFS scheduling strategy, assuming zero transfer overhead. Obtaining the smallest average energy by determining the optimal execution frequency of each generic block is the primary task of this ILP model. An extension to the ILP formula finds the optimal execution frequency and places it in the program to insert the conversion instructions; this helps with the DVFS conversion overhead. This method fails to address the issue of resource idleness for a longer period.

An indicator of energy awareness is the growth of studies focusing on efficient task scheduling in response to rising computer system energy consumption. [24], [25] discuss research on DVFS-based energy-efficient task scheduling. Meeting quality of service requirements while optimizing energy consumption is the objective of these task schedules. Assigning CPUs and scaling frequencies are the two typical stages. Assigning tasks to the processor in such a way that they are all complete by the due date while using the least amount of energy possible is the goal. Quality of service requirements for real-time tasks are expressed in [26] using a general model. They improve energy-aware task allocation and the ability to relocate tasks during runtime. Upon a task's completion during runtime, the remaining tasks could not execute to their full potential. To cut down on power usage, they suggested a method for local task relocation and tweaked the frequencies of the related CPUs. This method fails to take into account the time frames of the tasks assigned.

At the OS, hardware, virtualization, and data center levels, the authors of [27] mapped out a taxonomy of energy-efficient computer system design, and they also discussed the causes and problems of excessive power/energy utilization. They looked into several methods of limiting power usage from the operating system level using DVFS and other algorithms and strategies for power savings, and they sorted them all out. However, they failed to address the issue with optimal solutions. An Energy-Aware Task Scheduling Algorithm (ETSA) was suggested by Rohith et al. in [28] to tackle the issues that are directly tied to task consolidation and scheduling. After going through a normalization method, the suggested ETSA algorithm takes into account the task's overall resource utilization, and completion time, and uses that information to determine the best time to run the program. It fails to minimize the energy consumption to the maximum extent. Furthermore, the authors in [29] brought attention to the research challenges associated with the conflicting demands of improving the QoS provided by cloud services while simultaneously cutting down on the power usage of data center assets. In order to combine data center capabilities while lowering the influence on the quality of service objectives, they tackled the idea of designing an energy-efficient data center controller. For the purpose of conducting energy-efficient operations, they looked at methods of managing and coordinating data center resources. In addition, they suggested resource controller cooperation and proposed the idea of a central controller. Various mechanisms and frameworks for designing energy-effective data centers were discussed by the authors in [30]. Operating systems, virtual machines, and software applications were the subjects of their investigation into various power models.

An adaptive task-scheduling technique was proposed by Yao Sheng et al. [31]. An evolutionary algorithm was suggested (E-PAGA) to accomplish adaptive regulations for various energy and performance needs in cloud tasks after they modelled the energy of virtual machines for scheduling work in the cloud. In order to choose the next generation using distinct energy and performance measurements, they developed two distinct fitness criteria. They suggested customizing the cloud work to each user's needs by adaptively adjusting the target performance and energy. In practice, this method fails to uphold the deadlines of the tasks scheduled. In [32], the authors address scheduling issues with multiple purposes by using the improved PSO algorithm (AMTS). They begin by formulating the scheduling problem. Optimal resource utilization, average cost, average power consumption, and task completion time are achieved by advancing the task scheduling policy. The acceptance of the adaptive acceleration coefficient is contingent upon the preservation of particle diversity. This method fails to address the problem of achieving timelines successfully for all the assigned tasks.

Authors in [33] proposed Energy and Performance-Efficient Task Scheduling [EPETS] to solve the energy usage problem with deadline restrictions in heterogeneous virtualized cloud computing. The suggested method aims to provide good performance while reducing total energy usage within the given time limitations. In order to achieve the deadline with minimal energy consumption, the task reassignment algorithm assigns work to a machine with a medium or low processing speed with relative ease. However, these method incurs huge overhead and employs complex calculations. All the above methods discussed lack in addressing the energy consumption problem while maintaining the performance along with the timelines of all the tasks in a heterogeneous cloud environment.

This research article presents a novel proposed method called Energy Efficient Task Re-scheduling (EETRS) for processing jobs on heterogeneous computing virtual machines using various components promptly. A variety of virtual computers build the system, and the scheduler is responsible for allocating work according to the system's quality of service standards. In the simulation stage, experiments were conducted on the real test-bed traces to evaluate the efficiency and efficacy of the suggested scheme.

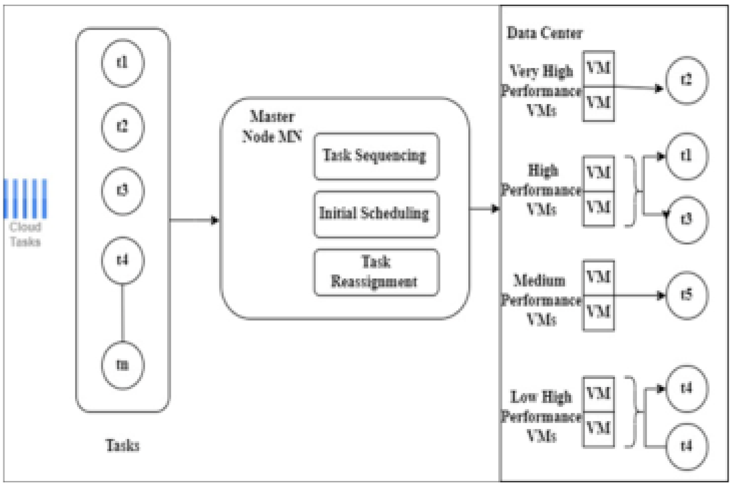


Fig. 3.1: Architectural framework of the proposed system

3. Proposed Method. The energy-efficient task scheduling problem on heterogeneous virtual machines is addressed in this article by proposing a method called Energy Efficient Task Re-scheduling (EETRS), which employs several strategies. Here are a few ways in which the EETRS algorithm differs from the aforementioned methods. The proposed method comprises the following steps:

1. An equitable trade-off between energy efficiency and task scheduling can be achieved by using the energy-efficient task priority framework, which is proposed in this paper.
2. The task finish time calculation is taken into account, which aids in enhancing system performance, decreasing energy usage, and preventing deadline breaches.
3. There are two parts to EETRS. The first step is to plan out all of the tasks that are needed to meet the SLA. The second step involves redistributing tasks to make the energy stability of the resources better

3.1. Proposed Architecture. Fig.3.1 shows the different components of the architecture that this study devises. Three components make up the suggested architecture. In the initial stage, users submit their independent tasks to the framework. The master node (MN) oversees multiple components in the management layer to ensure a successful execution in the second step of the proposed method. There are three main components: job reassignment, task sequencing, and initial scheduling. The research also takes into account the public data center single cloud, which includes four different kinds of heterogeneous virtual machines (VMs): small, medium, large, and very large. If you send tasks to the MN, the scheduling procedure will begin as follows and the proposed method will accomplish the following phases

1. Task sequencing is used to ease scheduling processes. These rules are applied to both submitted tasks and tasks that have come statically in the system.
2. To cut down on execution time and make deadlines, the scheduler initially tries to assign maximal and limited finish time tasks to the quicker machines.
3. To decrease energy usage without compromising deadlines, the faster machines should have their work redistributed to the slower ones

3.2. Task and Resource Features. A wide variety of independent tasks are now present in many large applications and businesses (e.g., manufacturing, disease diagnostics, etc.). Each task completes its own set of instructions. Different tuples of requirement quality (such as priority, workload, and due deadline) are stored for each job. However, there is a stringent standard for processing all tasks due to their characteristics. Due to this non-trivial challenge, Quality of Service (QoS) aware scheduling for distinct jobs in the cloud system has been taken into consideration. Under the QoS constraints, all jobs necessitate resource-intensive services and are computationally intensive. Amazon, Google, Alibaba, and Azure are just a few of the several public cloud providers that provide a range of service quality to their respective clientele. Virtual machines are used to access the services. Anyone may access these services anytime, anywhere because of the Internet. Computer systems and their ancillary components, including storage and telecommunications networks, are stored in data centers, which can be physical buildings, specific areas within larger buildings, or even clusters of buildings. A wide variety of services offered by the aforementioned suppliers allows the data center to host applications that require a lot of processing power and other resources. To efficiently conduct those mentioned compute-intensive operations, the data center can incorporate a variety of virtual machines into the system, such as a mix of vendors. Scheduling compute-intensive tasks in the cloud while considering their energy consumption and quality of service requirements is a complex and challenging issue. This research takes into account several virtual machines (e.g., Amazon) to manage the scheduling problem of many jobs. When compared to other sellers, Amazon's low-priced services are the key draw for customers.

3.3. Assumptions. During energy-efficient scheduling, this study takes into account the following assumptions about the problem. All jobs are distinct, with their workloads and due dates of personal ones.

1. With this setup, there is no need for communication between nodes because all tasks are scheduled on the same data center. Once tasks are assigned to virtual machines, they can be reassigned to other machines without affecting deadlines.
2. Since all machines share the same data center, only the power consumption of each machine is taken into account, which helps to reduce energy usage.
3. Task failure scenario was not considered in this study because it was assumed that all virtual machines could scale up and down. itemOn the other hand, the deadline requirement is considered when tasks are rejected.

3.4. Methodology. The process of the proposed method EETRS is detailed below:

1. The execution time of the task is calculated based on the size of the task and the computing power of the virtual machine.
2. The finish time of the virtual machine with current jobs is determined next.
3. The finish time of the task under consideration on a particular virtual machine is determined based on the finish time of the VM, as the third step.
4. Finally finish time of the task is compared with the deadline to assign it to that VM.
5. Then energy consumption of the VM is computed using its computing power and execution time taken by that task.
6. Finish time vs Deadline and Energy consumption of the VM is taken into consideration for reassigning the tasks for energy-efficient task scheduling.

To satisfy user needs, this study aims for a virtualized cloud with a collection of virtual machines ($V =$ v1, v2,..., vM). Different computing speeds along with powers create heterogeneous virtual machines, represented by Cj(j = 1,..., M) and Pj, respectively. The CPU power performance of VM vj is measured by MIPS (Million instructions per Second), as given in Equ.3.1

$$PW_j = \frac{C_j}{P_j} \tag{3.1}$$

where C_j is the Computing Speed and P_j is the Power

The suggested strategy EETRS aims to plan tasks to reduce cloud energy usage and meet data center VM deadlines. All jobs can be run on either faster or slower VMs. The tasks are delivered in a set of T= t1,t2,t3,...,tN and arrive simultaneously in a cloud system.

Each job ti contains parameters ti=si, di, where si(i = 1,...., N) and di represent task size and deadline. Initially, to preserve cloud resources, tasks are assigned to slower VMs to meet deadlines and reduce energy. The energy consumption of a task ti is determined by its power Pj and execution time EXTi, taking into account VM CPU processing power variations.

Let EXTfm i,j and EXTsm i,j represent task ti execution time on VM vj, where fm represents “faster machine” and sm represents “slower machine”. The execution time of task ti is calculated as stated in Equ.3.2.

Execution time of the task on a virtual machine j

$$EXTi = \frac{S_i}{C_j} \tag{3.2}$$

where Si is te Size of the Task ti, Cj is the Computing Speed of the Virtual Machine j, Finish time FTj, 0 is initialized to 0 for each VM vj.

The execution time of the current task tk as mentioned in Equ.3.3.

$$\sum_{k=1}^{N} x_j, kEXTk \tag{3.3}$$

The current task $t-k$ and the finish time of the preceding job $t_k - 1$ determine the finish time FTj,k of the virtual machine vj when task tk is being executed.

The finish time of the Virtual Machine is given in Equ.3.4.

Finish Time of the VM_j

$$FT_{j,k} = T_{jk} - 1 + \sum_{k=1}^{N} xj, kEXTk \tag{3.4}$$

The task's completion time on VM vj is described using FTfm j,k and FTsm j,k, respectively.

The completion finish time of task ti is expressed using Ffm I (for faster machine) and Fsm i (for slower machine). Starting with each virtual machine (VM),

The finish time of task ti on VM vj can be computed as follows using Equ.3.5. Finish time of the task ti

$$Fi = \sum_{j=1}^{M} x_{i,j} * FT_{j,k} \tag{3.5}$$

where $FT_{j,k}$ = Finish time of the virtual machine.

Finish time Fi of the Task Ti completion time should be less than or equal to the deadline di, as stated in Equ.3.6.

$$Fi \leq di \tag{3.6}$$

3.5. Energy consumption model. The central processing unit (CPU), random access memory (RAM), disc storage, and network interface (NIC) are the primary determinants of data center computing server energy usage. Furthermore, there are two types of energy consumption: static and dynamic [34]. A computer's central processing unit (CPU) and other dynamic components account for the bulk of its static energy usage. Thus, we focus our attention on dynamic energy use when we construct models of energy consumption.

The power Pj of the virtual machine (VM) vj and the execution time (EXTi) of the process determine the energy consumption (Ei,j) that the task ti consumes while running on the VM vj as shown in Equ.3.6.

Energy consumption of VM:

$$E_{i,j} = Pj * EXTi \tag{3.7}$$

where, Pj is the power of VM, EXTi is the Execution Time of task ti.

When the faster virtual machine (VM) vj is used to perform task ti, the energy consumption of the task is shown as in Equ.3.7.

Energy Consumption on a faster machine:

$$Efm_{i,j} = Pj * ETi \tag{3.8}$$

By contrast, when the slower VM is used, the energy consumption is shown as in Equ.3.8.

$$Esm_{i,j} = Pj * ETi \tag{3.9}$$

The goal is to reduce cloud resource energy usage by completing all jobs as calculated below in Equ.3.9.

$$\min Etotal = \sum_{i=1}^{N} \sum_{j=1}^{M} x_{i,j} * Pj * ETi \tag{3.10}$$

where $x_{i,j}$ is the Task sequenced, Pj is the power of VM, EXTi is the Execution Time of task ti.

The proposed system EETRS schedules tasks efficiently using the above methodology. This approach reduces cloud resource energy usage and boosts system efficiency while fulfilling deadlines. Slower machines can use less energy than faster ones. However, slower equipment can dramatically impair system efficiency and deadlines. Assigning all jobs to a slower machine deadline increases violations and system performance.

This proposed EETRS technique optimizes the performance by reassigning the tasks dynamically by considering the deadlines and energy usage.

3.6. Proposed Algorithm. The section provides an overview of the proposed system and how it addresses the problem of scheduling tasks in virtual machines in an energy-efficient manner across various environments. Energy Efficient Task Re-Scheduling, or EETRS, is an algorithm with two steps. First schedule to finish all jobs by their due date. This allows for a decrease in the total amount of time it takes to execute jobs. But the faster machines meant that the original plan used a lot of energy, even though it could attain fair efficiency. Second, the proposed novel task reassignment mechanism addresses this gap and lowers minEtotal's energy consumption. Figure 3.2 is a flow diagram depicting the entire EETRS job scheduling algorithm. The algorithm is described below (Algorithm 1).

Consider:

1. FiFM = Finish time of the task on a faster VM
2. FiSM = Finish time of the task on a slower VM
3. Ei FM = Energy consumption on faster VM
4. Ei SM = Energy consumption on slower VM
5. di = deadline of the task.

Then,

```
    If FiFM <= di and Ei FM <= Ei SM
    Then allocate the Task ti to a faster virtual machine
If FiFM <=di and Ei FM > Ei SM and FiSM <=di
    Then allocate the Task ti to a slower virtual machine
If FiSM <=di and Ei FM > Ei SM
    Then allocate the Task ti to a slower virtual machine
Else (in all other cases)
    allocate the Task ti to a slower virtual machine.
```

4. Simulation and results. Here, the performance of the proposed Energy Efficient Task Re-Scheduling (EETRS) algorithm is shown, and how well it works with a variety of randomly generated task counts. EETRS is compared quantitatively with three other methods that are already in use: EPETS [33], AMTS [32], and E-PAGA [31].

4.1. Simulation Settings. Four distinct virtual machines (VMs) built on Amazon Elastic Compute Cloud (EC2) are taken into consideration. Table.4.1 shows the different virtual machines (VMs), their configurations, processing speeds, and power consumption. Although the processing rates and powers vary, the number of cores in each virtual machine is set to 1. Python is used to implement all parameters and algorithms. The system running on Windows Server has 8 GB of RAM and an Intel (R) Core (TM) i7-9750H CPU with 2.60 GHz.

4.2. Experimental Findings.

4.2.1. Task Sequinning. First, all tasks are categorized according to the FCFS and EDF methods for sequencing, which stand for First Come, First Serve, and Early Deadliest First, respectively. Task sequencing is a prerequisite to scheduling tasks on a network of diverse virtual machines.

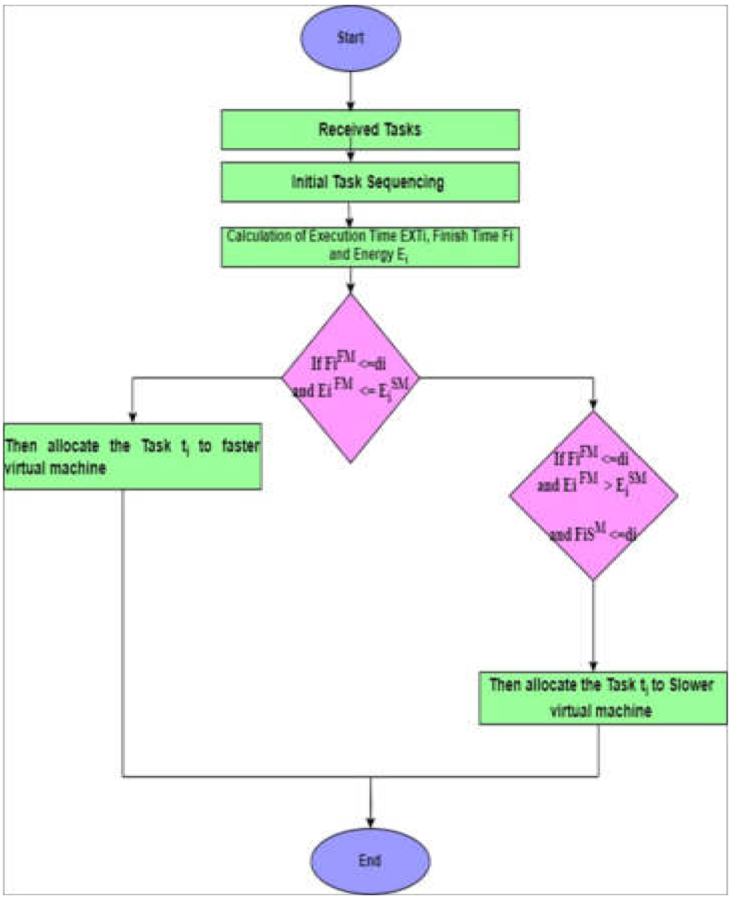

Start
Received Tasks
Initial Task Sequencing
Calculation of Execution Time EXTi, Finish Time Fi and Energy E_i
If Fi^{FM} <=di and Ei^{FM} <= E_i^{SM}
Then allocate the Task t_i to faster virtual machine
If Fi^{FM} <=di and Ei^{FM} > E_i^{SM} and FiS^{M} <=di
Then allocate the Task t_i to Slower virtual machine
End

Fig. 3.2: Flowchart of the Proposed methodology EETRS.

Table 4.1: VMs Specification

VM	Small VM1	Medium VM1	Large VM
Core	1	1	1
MIPS/Core	200	400	600
Power/Core	50W	100W	200W

Table 4.2: Initial Scheduling

Task	VM1 (Slower)	VM2 (Slower)	VM3 (Medium)	VM4 (Faster)
T1	yes			
T2		yes		

4.2.2. Initial Scheduling. The optimal energy efficiency of the cloud resources is not taken into consideration when all tasks are planned on separate machines in the initial stage, depending on the task series. The goal of this stage is to organize everything in such a way that the total execution time can be lowered while still achieving all of the deadlines. Initial scheduling is tabulated in Table.4.2.

Algorithm 10 Task Scheduling

```
Input: Ti (Task), Si (Task Size), Ci (Computation Power), EXTi (Execution Time), FTfm (Finish Time-Faster Machine), FTsm (Finish Time-Slower Machine), EiSM (Energy-Slower Machine), EiFM (Energy-Faster Machine), di (Deadline).
Output: Ti (Task Allocation to the Virtual Machine)
Begin
    Task Sequencing and Initial Scheduling
    For each Ti $\epsilon$ Qt do
    Execution Time, EXTi $\longleftarrow$ Si/Cj
    Finish Time of VM, FTj,k $\longleftarrow T_{j,k} - 1 + \sum_{k=1}^{N} xj, kEXTk$
    Finish Time of the Task Ti: $Fi \longleftarrow \sum_{j=1}^{M} x_{i,j} * FTj, k$
    Energy Consumption, $E_{i,j} \longleftarrow Pj * EXTi$
    Task Reassignment
    If FiFM <= di and Ei FM <= Ei SM
    Then allocate the Task ti to a faster virtual machine
    Else If FiFM <= di and Ei FM > Ei SM and FiSM <= di
    Then allocate the Task ti to a slower virtual machine
    Else If FiSM <= di and Ei FM > Ei SM
    Then allocate the Task ti to a slower virtual machine
    Else (in all other cases)
    Allocate the Task ti to a slower virtual machine
    end if
    End
```

Table 4.3: Task Re-Scheduling

Task	Execution Time	Deadline	Finish Time		Energy Consumption		Task Reassignment
T1	3	5	VM1S	6	VM1	150	T1 is assigned to VM2, due to less energy consumption and meeting deadline comparatively
	3	5	VM2S	5	VM2	150	
	2	5	VM3M	5	VM3	200	
	1	5	VM4F	4	VM4	200	
T2	4	6	VM1S	8	VM1	200	T2 is assigned to VM3, due to less energy consumption and meeting deadline comparatively.
	5	6	VM2S	8	VM2	250	
	2	6	VM3M	5	VM3	200	
	3	6	VM4F	4	VM4	600	

4.2.3. Task Re-Scheduling. The execution time, finish time and energy consumption are computed for the tasks sequenced and virtual machines available using the aforementioned equations. The stats are tabulated below in Table 4.3. Based on the calculations, tasks are reassigned to achieve energy efficiency while meeting the deadlines.

Task T1 is re-assigned to the slower virtual machine VM2 as execution time and deadline are within reachable limits and the energy consumption is less compared to other virtual machines. Task 2 is reassigned to Medium virtual machine VM3 due to less energy consumption and meeting the deadline compared to other virtual machines.

4.3. Performance Analysis. EETRS is compared to EPETS [33], AMTS [32], and E-PAGA [31] in terms of execution time, cost, energy usage, and resource utilization. The suggested method outperforms existing methods in all performance parameters, as shown below.

Table 4.4: Total Execution time compared with No. of Tasks

No. of Tasks	Proposed EETRS	EPETS	AMTS	EPAGA
10	616	620	812	1078
20	982	996	1134	1256
40	1274	1292	1354	1567
60	1656	1663	1838	2085
80	1864	1872	2092	2234
100	2005	2028	2384	2476

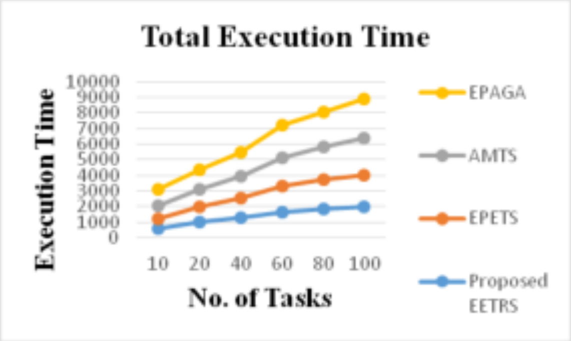


Fig. 4.1: Total Execution time vs No. of Tasks.

Table 4.5: Total Execution cost compared with No. of Tasks

No. of Tasks	Proposed EETRS	EPETS	AMTS	EPAGA
10	90012	90036	93245	97425
20	96054	96114	98423	102243
40	101224	102315	103245	107452
60	109018	110031	120234	123314
80	125423	128535	138354	145564
100	128434	133537	143637	156672

4.3.1. Total Execution Time. As seen in Fig.4.1 execution time was plotted versus job count. It is evident that EETRS takes equivalent execution times as other methods EPETS, AMTS, and EPAGA lag in this field a bit. Table.4.4 shows the total execution time of all the algorithms EETR, EPETS, AMTS, and EPAGA. From the graph in Fig.4.1, and Table.4.4, it is clear that the proposed EETRS shows less execution compared to EPETS, AMTS, and EPAGA algorithms. The proposed EETRS shows 1 % less execution time than EPETS over an average no of jobs.

4.3.2. Total Execution Cost. Fig.4.1 shows an association between job number and total execution cost. This graph shows that EETRS has a lower total cost than EPETS, AMTS, and EPAGA. Table.4.5 shows the total execution costs for the proposed EETRS and other existing algorithms. The stats show that EETRS exhibits a lower cost of 14% when compared to other algorithms over an average no of jobs.

4.3.3. Energy Consumption. The suggested EETRS algorithm has much lower energy usage compared to the EPAGA, EPETS, and AMTS algorithms. Fig.4.2 and Table.4.6 present the energy consumption comparison of proposed and existing algorithms. The benchmark algorithms generate massive energy consumption preparation, as shown by this figure and table. The rationale for this is that approximate task execution position generation does not yield optimal results when VM performance is reasonably high. The suggested approach

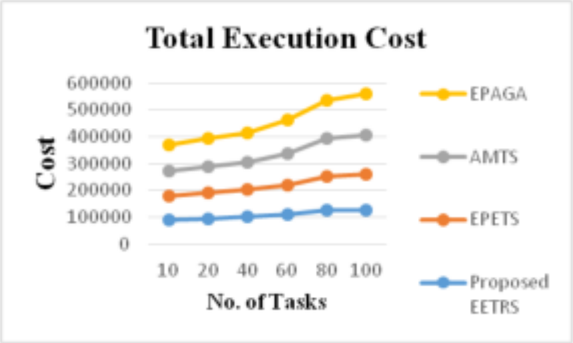


Fig. 4.2: Total Execution cost vs No. of Tasks

Table 4.6: Energy consumption compared with No. of Tasks

No. of Tasks	Proposed EETRS	EPETS	AMTS	EPAGA
10	4	7	9	12
20	6	9	12	17
40	8	13	16	22
60	12	17	21	27
80	14	21	24	32
100	17	24	27	38

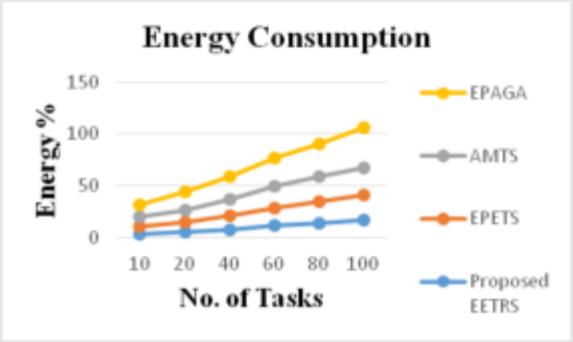


Fig. 4.3: Energy consumption vs No. of Tasks.

EETRS optimizes the use of slower computer resources, uses energy-efficient job sequencing, and reassigns tasks to machines with more spare time, resulting in higher energy efficiency than the existing algorithms. In addition, free machines can sit idle, significantly reducing energy consumption, when jobs on faster machines are transferred to slower machines. Because of such, the energy consumption outcome. The proposed EETRS shows a less energy consumption of 3% compared to other existing methods.

4.3.4. Average Resource Utilization. When it comes to scheduling tasks, the cloud data center's resource utilization is important. The utilization is contingent upon the workload execution that various users request. Equ.4.10 measures the average utilization of resources and performances.

$$AvgUtilization = \frac{\sum_{j=1,i=1}^{M} Si/Cj}{M} \tag{4.1}$$

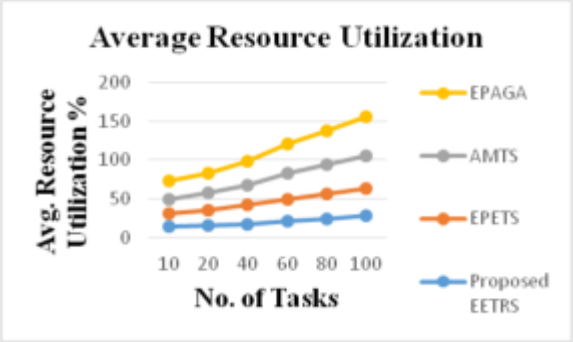


Fig. 4.4: Average resource utilization vs No. of Tasks

Table 4.7: Average resource utilization compared with No. of Tasks

No. of Tasks	Proposed EETRS	EPETS	AMTS	EPAGA
10	14	17	19	23
20	16	20	22	25
40	18	24	26	31
60	22	28	33	38
80	25	31	38	43
100	28	35	43	49

where Si is the Size of the task, Cj is the Computing power of VM.

When it comes to scheduling and accomplishments, the EETRS system improves resource consumption (e.g., CPU usage accurately without wastage). On the other hand, these resource and efficiency savings from system utilization in a network of heterogeneous virtual machines have not been considered in previous research. The multiple reasons why EETRS makes efficient use of its resources without wasting are listed below. (i) In the first stage, all tasks are planned to meet the needs of the SLA. (ii) In the second stage, tasks are redistributed to make the resource more energy stable. (iii) Deadline times save resources by task sequencing, thus no pool of resources is lacking. Thus, the suggested approach EETRS improves the system's overall reliability and maintains its full performance in the long run. Figure 4.4 and Table 4.7 show the average resource utilization of all the algorithms including proposed EETRS and existing under consideration. The proposed EETRs exhibit effective resource utilization with 3% lower when compared with other algorithms.

5. Conclusion. This article covers the Energy Efficient Task Re-Scheduling problem with deadline constraints in heterogeneous virtualized cloud computing. The proposed strategy EETRS performed better than all existing schemes, according to the simulation results. In order to meet the deadline, the proposed solution aims to achieve good performance while reducing the total usage of energy and improving the overall data center performance by lowering the execution time. The purpose of this study is to present an EETRS heuristic method that combines initial scheduling with task reassignment scheduling. Before any scheduling can take place, tasks must be sequenced, primary assignments made, and execution slots distributed. This method initially tried to assign the fastest machines maximum and lowest slack time tasks in the first scheduling without considering energy optimization. In the second stage, to reduce power consumption and still make the deadline, the task reassignment algorithm moves jobs from a fast computer to a medium or slower one. The suggested EETRS algorithm outperformed the existing methods in terms of total execution time, total execution cost, energy consumption, and resource utilization according to the simulation findings. In subsequent research, quantum-inspired methods can be included in a heterogeneous virtualized cloud environment to refine the proposed EETRS algorithm for scheduling tasks with minimal energy consumption.

REFERENCES

[1] Dayarathna, Miyuru, Yonggang Wen, and Rui Fan. ”Data center energy consumption modeling: A survey.” IEEE Communications surveys & tutorials 18.1 (2015): 732-794.

[2] Kliazovich, Dzmitry, et al. ”Energy consumption optimization in cloud data centers.” Cloud services, networking, and management (2015): 191-215.

[3] Zafar, Saima, Shafique Ahmad Chaudhry, and Sara Kiran. ”Adaptive trimtree: Green data center networks through resource consolidation, selective connectedness and energy proportional computing.” Energies 9.10 (2016): 797.

[4] Wu, Yu, et al. ”Green data center placement in optical cloud networks.” IEEE Transactions on Green Communications and Networking 1.3 (2017): 347-357.

[5] Ghafari, R., F. Hassani Kabutarkhani, and Najme Mansouri. ”Task scheduling algorithms for energy optimization in cloud environment: a comprehensive review.” Cluster Computing 25.2 (2022): 1035-1093.

[6] Zhu, Xia, Mehboob Hussain, and Xiaoping Li. ”Energy-efficient independent task scheduling in cloud computing.” Human Centered Computing: 4th International Conference, HCC 2018, Mérida, Mexico, December, 5–7, 2018, Revised Selected Papers 4. Springer International Publishing, 2019.

[7] Chen, Huangke, et al. ”EONS: minimizing energy consumption for executing real-time workflows in virtualized cloud data centers.” 2016 45th International Conference on Parallel Processing Workshops (ICPPW). IEEE, 2016.

[8] Jena, R. K. ”Energy efficient task scheduling in cloud environment.” Energy Procedia 141 (2017): 222-227.

[9] Medara, Rambabu, and Ravi Shankar Singh. ”Energy efficient and reliability aware workflow task scheduling in cloud environment.” Wireless Personal Communications 119.2 (2021): 1301-1320.

[10] Panda, Sanjaya K., and Prasanta K. Jana. ”An energy-efficient task scheduling algorithm for heterogeneous cloud computing systems.” Cluster Computing 22.2 (2019): 509-527..

[11] Kak, Sanna Mehraj, Parul Agarwal, and M. Afshar Alam. ”Task scheduling techniques for energy efficiency in the cloud.” EAI Endorsed Transactions on Energy Web 9.39 (2022): e6-e6.

[12] Tang, Chaogang, et al. ”Energy-aware task scheduling in mobile cloud computing.” Distributed and Parallel Databases 36 (2018): 529-553.

[13] BEN ALLA, Said, et al. ”An efficient energy-aware tasks scheduling with deadline-constrained in cloud computing.” Computers 8.2 (2019): 46.

[14] Walia, Navpreet Kaur, et al. ”An energy-efficient hybrid scheduling algorithm for task scheduling in the cloud computing environments.” IEEE Access 9 (2021): 117325-117337.

[15] Yang, Jiachen, et al. ”A task scheduling algorithm considering game theory designed for energy management in cloud computing.” Future Generation computer systems 105 (2020): 985-992.

[16] Li, Xiaoping, et al. ”Energy-aware cloud workflow applications scheduling with geo-distributed data.” IEEE Transactions on Services Computing 15.2 (2020): 891-903.

[17] Lee, Heecheon, et al. ”Energy-efficient design of a novel double annular separation column using pinch pressure.” Industrial & Engineering Chemistry Research 59.32 (2020): 14398-14409.

[18] Sheikh, Hafiz Fahad, Ishfaq Ahmad, and Sheheryar Ali Arshad. ”Performance, energy, and temperature enabled task scheduling using evolutionary techniques.” Sustainable Computing: Informatics and Systems 22 (2019): 272-286.

[19] Ashraf, Adnan, and Ivan Porres. ”Multi-objective dynamic virtual machine consolidation in the cloud using ant colony system.” International Journal of Parallel, Emergent and Distributed Systems 33.1 (2018): 103-120.

[20] Aryania, Azra, Hadi S. Aghdasi, and Leyli Mohammad Khanli. ”Energy-aware virtual machine consolidation algorithm based on ant colony system.” Journal of Grid Computing 16 (2018): 477-491.

[21] Zhang, Qing, et al. ”Energy-aware scheduling in edge computing based on energy internet.” IEEE access 8 (2020): 229052-229065.

[22] Shu, Zhaogang, et al. ”Cloud-integrated cyber-physical systems for complex industrial applications.” Mobile Networks and Applications 21 (2016): 865-878.

[23] Qin, Yang, et al. ”Energy-efficient intra-task DVFS scheduling using linear programming formulation.” Ieee Access 7 (2019): 30536-30547.

[24] Wang, Songyun, et al. ”A DVFS based energy-efficient tasks scheduling in a data center.” Ieee Access 5 (2017): 13090-13102.

[25] Yu, Ke, et al. ”An improved DVFS algorithm for energy-efficient real-time task scheduling.” 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2020.

[26] Ahn, Gilseung, and Sun Hur. ”Multiobjective Real-Time Scheduling of Tasks in Cloud Manufacturing with Genetic Algorithm.” Mathematical Problems in Engineering 2021.1 (2021): 1305849.

[27] Tian, Wenhong, et al. ”On minimizing total energy consumption in the scheduling of virtual machine reservations.” Journal of Network and Computer Applications 113 (2018): 64-74.

[28] Zambre, Rohit, et al. ”Breaking band: A breakdown of high-performance communication.” Proceedings of the 48th International Conference on Parallel Processing. 2019.

[29] Akintoye, Samson Busuyi, and Antoine Bagula. ”Improving quality-of-service in cloud/fog computing through efficient resource allocation.” Sensors 19.6 (2019): 1267.

[30] Ahmed, Kazi Main Uddin, Math HJ Bollen, and Manuel Alvarez. ”A review of data centers energy consumption and reliability modeling.” IEEE access 9 (2021): 152536-152563.

[31] Shen, Yao, et al. ”Adaptive task scheduling strategy in cloud: when energy consumption meets performance guarantee.” World Wide Web 20 (2017): 155-173.

[32] Mubeen, Aroosa, et al. ”Alts: An adaptive load balanced task scheduling approach for cloud computing.” Processes 9.9 (2021):

1514.

[33] Hussain, Mehboob, et al. ”Energy and performance-efficient task scheduling in heterogeneous virtualized cloud computing.” Sustainable Computing: Informatics and Systems 30 (2021): 100517.

[34] Katal, Avita, Susheela Dahiya, and Tanupriya Choudhury. ”Energy efficiency in cloud computing data centers: a survey on software technologies.” Cluster Computing 26.3 (2023): 1845-1875.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jul 22, 2024
Accepted: Sep 26, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1180–1195, DOI 10.12694/scpe.v26i3.4381

APPLICATION OF BIG DATA ANALYSIS IN INTELLIGENT INDUSTRIAL DESIGN USING SCALABLE COMPUTATIONAL MODEL

ZHE ZHANG[*]AND HESHUAI ZHANG [†]

Abstract. Smart industrial design's incorporation of big data analytics is reshaping the manufacturing industry by boosting product innovation, optimizing design processes, and increasing overall efficiency. Massive amounts of data may be processed using scalable computing, leading to crucial insights that propel more informed, data-driven design choices. Integrating advanced analytics into current design workflows, dealing with diverse and large-volume data, and guaranteeing data quality and integrity are all obstacles to implementing extensive data analysis in industrial design. Many challenges must be solved, including keeping data secure and meeting the computational needs of real-time processing data. Intelligent industrial design benefits significantly from scalable computing's extensive data analysis capabilities, which allow systems to analyze huge quantities of data in real time. Its dynamic resource allocation achieves efficient resource utilization and optimum performance, guaranteeing that processing power scales with the demand. This research suggests an Integrated Concentric Framework for Intelligent Industrial Design (ICF-IID) that applies big data analysis using scalable computational resources. The framework analyses big datasets from different parts of the design and production process using powerful visualization tools, machine learning algorithms, and predictive analytics. Adaptive algorithms developed for unique demands in industrial design, strong data management protocols, and a distributed computing architecture for efficient data processing are essential components. The framework is useful for predictive maintenance, product lifecycle management, and design parameter optimization in industrial design. The framework may find design defects, predict equipment failures, and suggest improvements by analyzing historical and real-time data. The efficacy and scalability of the suggested framework are assessed through simulation analysis. These results show that it can efficiently and accurately process industrial data on a wide scale. Based on the results, the framework seems useful for making decisions in complicated design contexts and providing practical insights. The proposed method increases the efficiency ratio of 9.21%. accuracy ratio of 98.32%, product innovation ratio of 97.65%, scalability ratio of 97.41%, and optimized design process ratio of 96.21% compared to existing methods.

Key words: Big Data, Analysis, Intelligent, Industrial, Design, Scalable Computational Model, Integrated, Concentric.

1. Introduction. The industrial sector generates enormous amounts of large data in the age of big data, and this data has ultra-high dimensions [1]. It is a difficult task to handle this ultra-high dimension data, realize its potential, and create a data flow model appropriate for the modern production setting [2]. Currently, the intelligent industry design will profit more optimally from big data-driven analysis with the reciprocal assistance of associated developing technologies against the backdrop of Industry 4.0 [3]. The goal of the data analysis procedure is to increase design effects and management transparency [4]. According to the internal organization of the business, industrial design based on big data-driven analysis enhances the operation of the whole production system [5]. To maximize its financial gains, it efficiently utilizes design production resources by ICF-IID [6].

Intelligent techniques may be used to extract and refine valuable information from data, which is crucial in several areas like as product design, scheduling, prognosis and health management, and quality management [7]. Several obstacles persist in smart manufacturing systems owing to their intricate design and demanding performance criteria [8]. This special issue is dedicated to exploring scientific paradigms, models, techniques, and technologies that have a strong theoretical foundation and practical significance in reshaping big data analytics in the manufacturing industry [9]. The main emphasis of this paper is on intelligence methodologies and applications for big data analytics in industrial design [10]. These days, business management and industrial design are very much intertwined. The enterprise's operations and outcomes are directly affected by the pros

[*]College of Traffic Engineering, Shijiazhuang Vocational College of Finance & Economics, Shijiazhang, 050061, Hebei, China (hbkdzhangzhe@126.com)

[†]College of Traffic Engineering, Shijiazhuang Vocational College of Finance & Economics, Shijiazhang, 050061, Hebei, China (zhangheshuai2008@163.com)

and cons of design management [11]. An important part of accomplishing corporate goal management is design management. In the design sector, it's crucial for the all-encompassing administration of the center's many components [12].

Design management encompasses a wide range of activities that are directly tied to enterprise decision-making [13]. These activities include, but are not limited to, design planning, schedule planning, personnel management of designers, education, and departmental coordination to effectively convey the enterprise's purpose, culture, and management policy through design [14]. The enterprise's official operation of a real-time linkage industrial design information system software symbolizes the fundamental maturity of industrial design informatization [15]. However, there are still some procedures to fulfil before the system can be fully maintained and expanded independently [16]. A real-time connection big data analysis in industrial design information system using scalable computational approaches has arrived, and it's very advantageous in terms of scalability and simplicity of maintenance [17]. The maintenance cost is little, the workload is light, and even non-professionals like designers can do a good job at it on their own with just a little instruction with the help of ICF-IID [18].

The main objective of this paper is as follows:

∗ To improve product innovation and design processes using big data analytics. The framework aims to handle enormous data sets to provide critical insights that result in better, data-driven design decisions using scalable computing and sophisticated analytics.

∗ The ICF-IID framework efficiently processes data and combines robust visualization capabilities by developing adaptive algorithms, robust data management protocols, and a distributed computing architecture.

∗ To provide useful insights for decision-making in complicated design settings by efficiently and correctly processing large-scale industrial data, predicting equipment failures, finding design errors, suggesting changes, etc.

The remaining of this paper is structured as follows: In section 2, the related research work of intelligent industrial design is studied. In section 3, the proposed methodology of ICF-IID is explained and in section 4, the efficiency of ICF-IID is discussed and analysed.

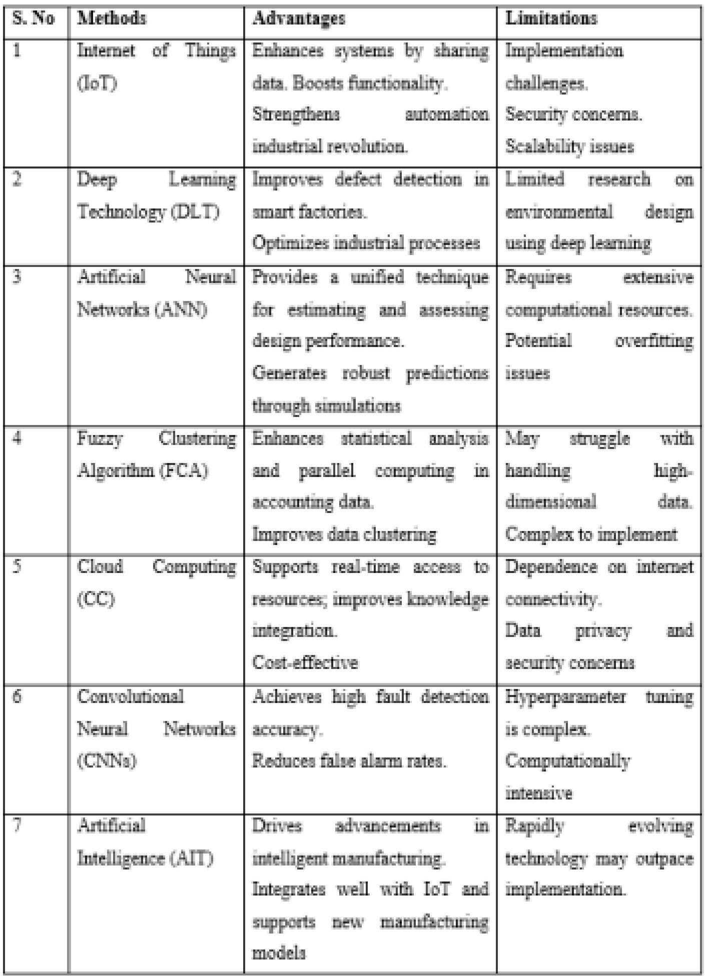
2. Related Studies. Industrial design is an approach to professional conduct that uses market research as a compass and addresses the human-product coordination connection in depth; its research is centered on the entire creation of products. Industrial design is an all-encompassing behaviour for product development that considers market demands, laws and regulations, economic considerations, etc., in addition to the four main points of the American Outstanding Idea Award: design innovation, aesthetic expression, environmental protection, and the degree to which manufacturers and users benefit.

Internet of Things (IoT). Connected sensors, devices, and services constitute what is known as the IoT. This network aims to enhance associated systems by sharing data and information over the Internet. By lowering prices, boosting functionality, expanding access to resources, and strengthening automation, the technologies linked to the IoT have greatly enhanced the quality of several current applications. Industries' embrace of the Internet of Things has sparked the fourth industrial revolution by Ahmed, S. T. et al., [19]. More and more, the emergence of the Industrial IoT holds the promise of better industrial management, optimized processes, and safer workers. Nevertheless, there are several big problems with the Internet of Things implementation that make it impossible to realize the full potential of Industry.

Deep Learning Technology (DLT). The goal of this project is to find the best design by developing a unified technique that uses simulation and ANN to estimate the functions of design parameters and assess the designs' performance by Chan, W. L. et al., [21]. This goal may be achieved by creating an integrated ANN technique. This approach uses the simulation to generate training instances for ANNs, which are then utilized to forecast the design's performance. The presentation also includes the methodology's structure and implementation approach in terms of both estimate and assessment of the design, the results demonstrate that the established technique works admirably.

Fuzzy Clustering Algorithm (FCA). With the help of cloud computing, it built a modern accounting data analysis platform. To make clustering of modern accounting data Ting, W. et al., [22] used the FCA. This improved the capacity for statistical analysis and parallel computing. Accounting data's statistical analysis capabilities and parallel computing efficiency are both enhanced by the intelligent data analysis platform,

Table 2.1: Summary of the existing methods



S. No	Methods	Advantages	Limitations
1	Internet of Things (IoT)	Enhances systems by sharing data. Boosts functionality. Strengthens automation industrial revolution.	Implementation challenges. Security concerns. Scalability issues
2	Deep Learning Technology (DLT)	Improves defect detection in smart factories. Optimizes industrial processes	Limited research on environmental design using deep learning
3	Artificial Neural Networks (ANN)	Provides a unified technique for estimating and assessing design performance. Generates robust predictions through simulations	Requires extensive computational resources. Potential overfitting issues
4	Fuzzy Clustering Algorithm (FCA)	Enhances statistical analysis and parallel computing in accounting data. Improves data clustering	May struggle with handling high-dimensional data. Complex to implement
5	Cloud Computing (CC)	Supports real-time access to resources; improves knowledge integration. Cost-effective	Dependence on internet connectivity. Data privacy and security concerns
6	Convolutional Neural Networks (CNNs)	Achieves high fault detection accuracy. Reduces false alarm rates.	Hyperparameter tuning is complex. Computationally intensive
7	Artificial Intelligence (AIT)	Drives advancements in intelligent manufacturing. Integrates well with IoT and supports new manufacturing models	Rapidly evolving technology may outpace implementation.

according to the simulation findings.

Cloud Computing (CC). Bohlouli, M. et al., [23] uses CC infrastructure to focuses on the idea of continuous research in delivering a knowledge integration service for collaborative product design and development. This article explains how cloud computing may help with knowledge integration as a service by offering features like knowledge mapping, merging, searching, and transferring in the product design process. Users are supported by the proposed knowledge integration services, which provide real-time access to resources for knowledge. Accessibility, efficiency, reduced costs, shorter time to result, and scalability are some of the benefits of the framework.

Convolutional Neural Networks (CNNs). Hyperparameter settings for CNNs and how they affect the reliability of fault detection findings. CNN is a revolutionary advancement in image processing that eliminates the need for human intervention or specialized process expertise. Instead, it uses hierarchical learning algorithms to autonomously produce robust features from large datasets of training data. Applying the suggested strategy yields minimal false alarm rates and great results for fault identification by Weimer, D. et al., [24].

Artificial Intelligence Technology (AIT). Li, B. H. et al., [25] assess the new era of 'Internet plus AI' and its fast-paced core technology development considering studies into AI's recent industrial applications. This era is causing a sea change in the manufacturing industry's models, means, and ecosystems, and it is also driving advancements in AI. Based on the integration of AI technology with information communications, manufacturing, and associated product technology, then suggest new models, methods, and forms of intelligent manufacturing, intelligent manufacturing system architecture, and intelligent manufacturing technology system. Table 1 shows the summary of the existing methods.

An industrial revolution is underway, propelled by the convergence of IoT, DLT, ANN, FCA, CC, CNNs, and AI. All of these technologies work together to make better decisions, analytics, and data processing [27]. Evidence from both theoretical and practical research shows that they are useful for analyzing data in real-time, finding design optimization opportunities, and detecting defects. Industries may boost accuracy, efficiency, and scalability by using these cutting-edge approaches. This will encourage innovation and help them stay ahead in the ever-changing technology market [28].

3. Proposed Method. The use of data-driven models is vital in modern manufacturing for managing the enormous amounts of information it generates. An all-inclusive approach involves processes such as collecting, storing, scrubbing, integrating, analyzing, mining and visualizing data. This paradigm promotes intelligent manufacturing since big data analysis supported by real-time dynamic perception leads to accurate decision-making. A parallelized k-means clustering method utilized for real-time data classification and analysis might be an adaptive algorithm in the ICF-IID framework. Here, the algorithm groups massive industrial process information into smaller ones according to shared characteristics, including product specs or production metrics. It learns from fresh data by constantly recalculating cluster centroids to see patterns and modify spontaneously. The ability of the framework to incorporate current data insights is essential for intelligent industrial design because it allows for decision-making that synchronizes with changes in operations that occur in real-time. It introduces a data-driven model to deal with the huge amount of information that is produced in manufacturing process. The technique encompasses data gathering, archiving, cleansing, integration, analysis, mining and visualization. In this case correct decision making in production setting's through big data driven analysis and real-time dynamic perception establishes a new model for intelligent manufacturing.

Speeding up distributed data processing among the suppliers and consumers improves the speed at which they communicate while emphasizing on gaining standardized interaction of Data. The first step towards industrial production's decision making process is creation of an exhaustive database alternatives from which choices can be made. Bill of materials, demand list, production planning, manufacturing job management, and other similar datasets comprise the industrial decision database's input data for decision tasks. Manufacturing businesses may gain a competitive edge by making more informed decisions with the use of verification standards and larger decision databases on industrial big data platforms. Before properly coupling simulation and optimization, intelligent manufacturing systems determine the important production characteristics to create a plausible production plan. Digital twins, which are essentially digital representations of physical objects, allow for optimization of simulations by reflecting the current state of physical operation in real time and by conducting simulation operations on virtual duplicates to generate fresh ideas is shown in figure 3.1.

$$F = \begin{pmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & & \vdots \\ y_{22} & \cdots & y_{2k} \end{pmatrix} + \begin{pmatrix} c_{11} & \cdots & y_{1k} \\ \vdots & & \vdots \\ c_{22} & \cdots & y_{2k} \end{pmatrix} \tag{3.1}$$

where y stands for variable design qualities $y_2 2$ and c for restrictions or conditions $c_2 1$, the equation 3.1 matrix form encompasses the many interconnected aspects of the design process $y_2 k$. The combined effect of design features and limitations is shown by adding up these matrices $y_1 k$, which emphasizes their collaboration F and overall influence on the design output.

$$m_L = \pi(y_p + r_s) + Z_m = \propto (w_k[j_{v-1}, z_x] - t_r) \tag{3.2}$$

The goal of optimization metric for the design, denoted by equation 3.2, m_L, is affected by the sum of the design parameter y_p and the scaling factor r_s, with the adjustment made by π. Incorporating an additional adjustment factor denoted as Z_m into the design process, this term is derived from $\propto a$ proportionality constant—and is

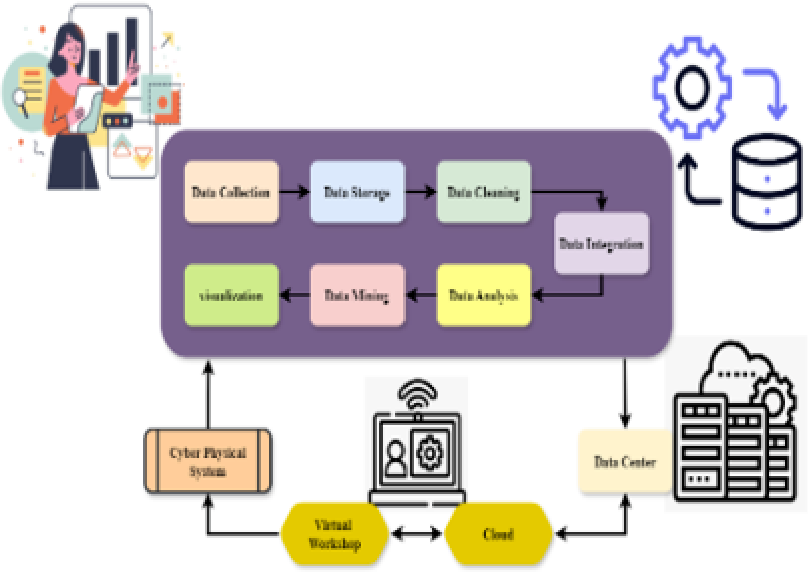


Fig. 3.1: Big data-driven intelligent decision making in industrial design

multiplied by the weighted difference within the measured variables $w_k[j_(v-1), z_x]$ and $t_r a$ temporal factor—to reflect real-time and historical data.

$$W_q = \partial(\propto (n - f_{q+1}) + k_w) = \forall^{mq}_{c-d}([j_{p-1}, z_k] + d_f) \tag{3.3}$$

The scale factor $\propto_v$, the constant n, the function $f_(q+1)$ representing the next design repetition, and another constant k_w are all components of the function that yields the weight or influence factor W_q. In a range $\forall^{mq}_{c-d}$, this weight represents the impact of previous iterations j_{p-1}, z_k and an extra factor d_f, and it is equal to the universal quantification (∀).

$$\partial_{q-p} = \forall(m_{w+1}q_{s+pq}) - S_{g+1} = \delta(Z_q[j_{p-1}, a_p] + d_1 q) \tag{3.4}$$

Except for a scaling factor S_{g+1}, the equation 3.3, $m_{w+1}q_{s+pq}$ captures the overall impact of incremental modifications ∂_{q-p} and their interaction with designer parameters $Z_q[j_{p-1}, a_p]$. Applying $d_1 q$ a proportionality constantto the impact of the variable Z_q classified by prior iteration and an adjusted factor equals this differential.

$$\min \sum_{i=1}^{q} (\nabla_m + \forall_{dq} - 1), \;\; p, w + 1C_v > hj_{k+1} + lqw - (m + sgt) \tag{3.5}$$

In the previous iteration, the design parameter $\forall_d q$ had a universal impact, and the equation 3.4, ∇_m represents a gradient of a design metric. The design constraint C_v is tested against a function including hj_{k+1}, lqw, and other variables in the conditional component $m + sgt$ representing additional modifying factors.

The incorporation of Big Data into smart manufacturing is seen in figure 3.2. The gathering of accessible big data from places like sensor data and log files is crucial to this. All these places add to the mountain of data that smart manufacturing processes need. Smart manufacturing adopts big-data techniques to address major issues involving traffic reduction and optimal timeframes for increased efficiency in operations and flow of information. Saving Big Data is an important part of this process as well as preserving relevant & significant ones. For effective use of Big Data applications in smart manufacturing there are certain requirements or criteria that must be met like Completeness and Correctness when gathering Big Data from different sources. To ensure

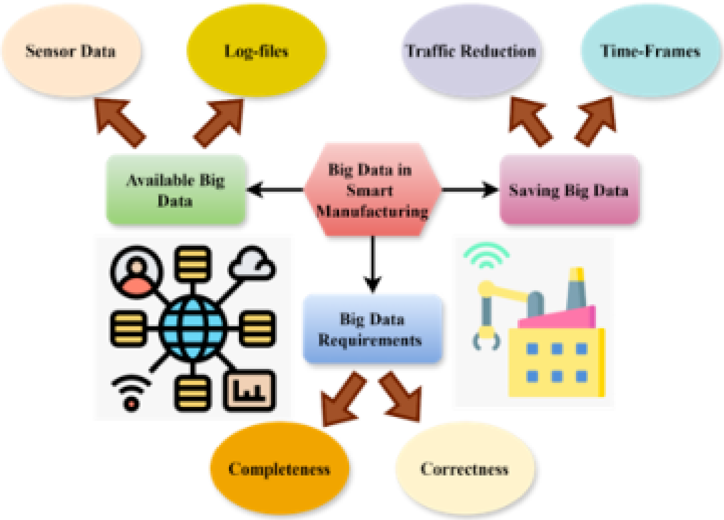


Fig. 3.2: Big data in smart manufacturing

completeness one has to ensure that he or she captures all possible required variables whereas correctness means coming up with reliable results from accurate input values only thus excluding any form of errors or mistakes due to inaccurate computations being used. Smart manufacturing uses Big Data for the precise collection facts about traffic, management traffic, efficient storage.

$$F_r = \sum_{i=1}^{q} T_l + \propto_k - \sqrt{\frac{1}{s} + \sum_{k=1}^{Q}(e_k + em_1)} - \frac{P_{correct}}{Q} + (1-k) \tag{3.6}$$

The variables $T_l + \propto_k$ reflect a complicated error-related component containing individual error terms e_k and an extra error modifier em_1, while the time or task-based factor T_l and the scaling constant $p_correct/Q$ are represented by the equation , $(1-k)$. In addition to a linear term $\propto_k$, the equation includes an adjustment factor $(e_k + em_1)$ that represents the accuracy of forecasts or corrections over F_r occurrences.

$$M(k) = -\sum_{i=1}^{q} s(jk) + \log e_s(z1 - z2) - 1 + \sum_{k=1}^{e} ks + 1w \tag{3.7}$$

Equation 3.7 sums up the cumulative effect of design parameters $M(k)$ across $s(jk)$ iterations, with the impact of particular design changes $(z1-z2)$ accounted for by the logarithmic term $loge_s(z1-z2)$. The scaling factors ks and the incremental factor $1w$ are included in the equation, along with constant adjustments and an extra summation $k = 1$. To maximize the total design metric $M(k)$, this formulation takes into account several factors.

$$f(b,w) = \sqrt{\sum_{k=1}^{q}(fs + pj) - (1 + mt)} - \frac{1}{1 - g - z^1} + (kp - wa) \tag{3.8}$$

The effects of factors fs and pj, modified by constants 1 and mt across q iterations are aggregated in the square root term $\frac{1}{1-g-z^1}$. An extra scaling and weighing factor are included in the linear term $(kp - wa)$, while a non-linear adjustment is introduced by the fraction.

$$vef(m,k,we) = qs(n+1) + phj(F,kp) - g^{-jpy} + hjv_p \tag{3.9}$$

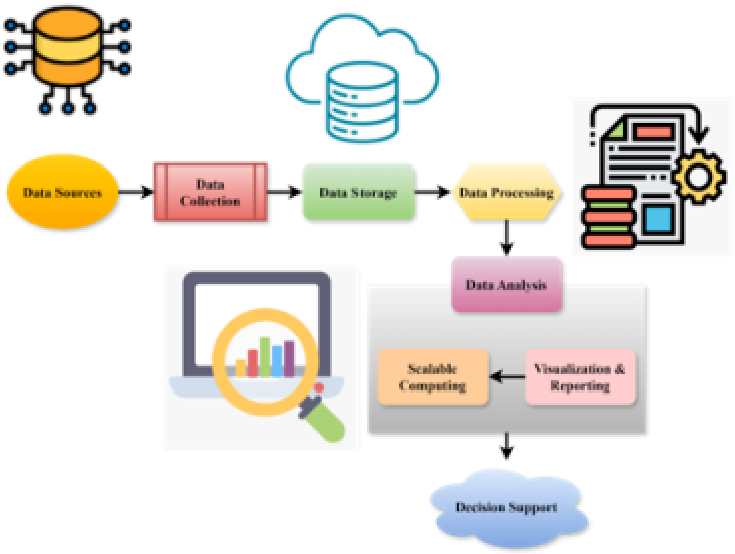


Fig. 3.3: Proposed method of ICF-IID

The exponential function $phj(F, kp)$ is reliant on parameters F and kp, whereas the scaling factor qs applied to an increased variable $(n + 1)$ is represented by $qs(n + 1)$. An exponentially decaying component affected introduced by the term g^{-jpy}, and an alteration based on another design component hjv_{-p} indexed.

Data Sources are at the beginning of the data lifecycle diagram showing how it moves through several phases in a structured system aiding in decision-making/choice-process/participation/selections/determinations. The first step is to collect data from multiple sources, making sure to get the right information. After collection, the data is sent to Data Storage and kept in a secure database for easy management and retrieval purposes. At this stage, it involves transforming raw data into another form that can be used in future analysis purposes. Making sure that such information has been properly cleaned and organized, prepares it well for further analysis. Following the data processing in Figure 3.3, there is a subsequent examination or study with results. This helps identify trends, correlations and patterns within the studied dataset. Such findings go to scalable computing, visualization, and reporting, where they are made less complicated for understanding (scalability) and simultaneously thoroughly explain all-important aspects of these reports. The Decision Support gets enhanced by such reports which add value to guiding both strategic and operational decisions on useful insights gained through efficiency of big data utilization in smart manufacturing processes; hence providing good results for decision making process towards improving efficiency and effectiveness of data-driven workplaces. Using this structured approach, which entails a continuous flow of data from collection through decision-making, ensures effectiveness and efficiency in data-driven workplaces.

$$U_{kp} = \frac{\sum_h^s j + pw}{2}\sqrt{1 - pu} + \Big(\sum_u^{g+fg} ng\Big) - (ks - pw) \tag{3.10}$$

The average impact of parameters $\sum_h^s j + pw$ iterations, modified by a factor involving kpu, is computed using ks-pw. The second term, $\sum_h^s j + pw$, denotes the total impact of ng across a wider range of values $g + fg$, and the last term, $1 - kpu$, takes into consideration further adjustments for scaling and weighting. Optimizing the overall model metric U_kp, this equation balances several impacts for increased design efficiency by integrating these aspects.

$$k(a_{s,w} = z_{w,q}|y_2) = \frac{fsw(r_f, gp)}{\sum_{l=1}^{W}(k_q + 1)} - fst(1 + phj) \tag{3.11}$$

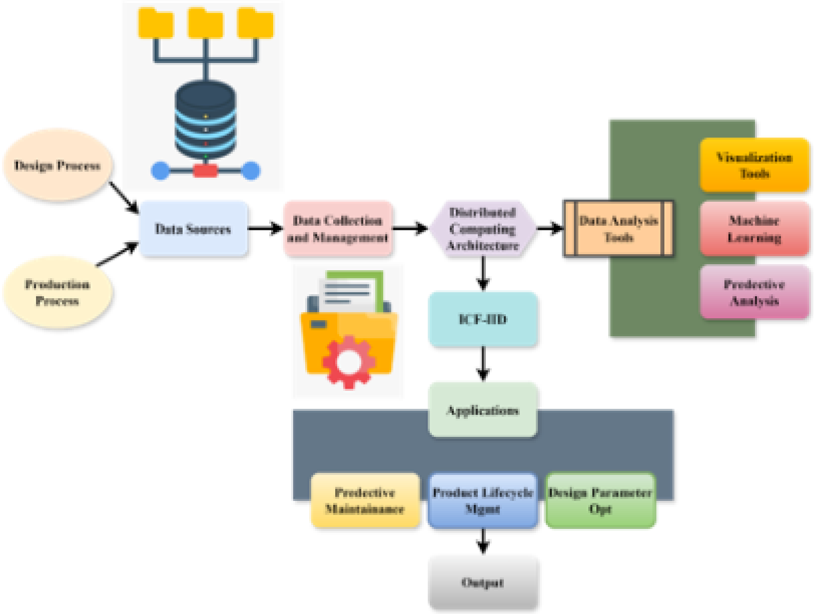


Fig. 3.4: Process flow of integrated concentric framework for intelligent industrial design

Within the context of $z_{w,q}$, the equation 3.11, $z_{w,q}$, denotes a conditional term that is reliant on $a_{s,w}$. The complex function $fsw(r_f, gp)$ may be computed by $\sum_{l=1}^{W}(k_q+1)$, while the iterative corrections can be seen in the denominator, which accumulates over W in terms of $fst(1+phj)$.

$$HC = h_k + W_{K+pk} - De = D^u_{f+1} - (F_{h+jk} + V_m(l+1)) - (q_w + y_q) \tag{3.12}$$

The equation 3.12 shows the additive parts that go into HC, and De is the element that subtracts from it. While F_{h+jk} subtracts the combined impact of $V_m(l+1)$, the. Finally, the design process is reflected when $q_w + y_q$ makes more modifications.

$$N(\forall + \propto) = +\sum_{v=1}^{Q} \log r(m^q|av_{p-1}) - Z_a(y_z + prst) - \partial_{q+1}(j+kt) \tag{3.13}$$

The variables that affect are represented by the equations $N(\forall + \propto)$. The amount being calculated is $(m^q|av_{p-1})$. The sum of the logarithmic evaluation of complex equations involving $Z_a(y_z + prst)$, with dependents on $\partial_{q+1}(j+kt)$.

$$r = 1 - \frac{\sum_x^f m2 + G - K}{p - q^3} + \frac{1}{12} - \sum_{b=1}^{2}(C_l^4 + D_m) - (rsg - mkp) \tag{3.14}$$

A complex ratio modified by p and q^3 is computed using the equation 3.14 $\sum_x^f m2 + G - K$, which influences the main term $C_l^4 + D_m$ at the beginning of the equation. The constant factor is represented by $(rsg - mkp)$.

The figure 3.4 shows how data is used in industry design and manufacturing process thus giving a good foundation. The starting point for all these processes are data sources which encompass the entire design and production processes. Data collection and management takes these inputs into account, with an eye on maintaining the accuracy and reliability of the collected information. Once data has been collected, it is processed using a distributed computing architecture. This design stresses efficient processing and scalable computing so as to efficiently manage massive amounts of data. The next step involves using data analysis tools such as visualization tools, machine learning, predictive analytics etc., to examine the processed data. These tools help in pattern recognition so that can make some educated presumptions. Through adaptive algorithms and strong data management, the ICF-IID incorporates analytic insights into its design processes. Predicative maintenance; Product lifecycle management; design parameter optimization are some of framework uses that results in useful output at end of it all. The smooth transition from practical implementations to industrial designs and manufacturing efficiency optimizations made possible by this systematic approach as depicted by figure 3.4.

$$U_m = \Pi^S_{p\partial R}(c_f + \propto_q - rs) + \Pi^T_{p\propto W}(b_n - U_{y+1}) + \frac{1}{4}(a(e) + \propto_{1-q} (m)) \tag{3.15}$$

In the intelligent designing framework, the metric U_m is defined by equation 15. The design process is encapsulated by the iterative adjustments and complexity of $c_f + \propto_q - rs$, which is a product of equations or transformations involving $(b_n - U_(y+1))$. Their respective products illustrate additive contributions and iterative dependencies via the expressions $(a(e) + \propto_{1-q} (m))$.

$$\lim_{n\to\infty}(1 + \frac{1}{w})^{pk} = fms_{ew} + Hwp(l_q, v_{rst}) + \log e_p(k+1) \tag{3.16}$$

The parameters pk determine the exponential growth of the equation 16, $1 + 1/w$. This exponential growth rate is estimated to be $fms_e w$ according to the equation, which is probably a function involving elements $Hwp(l_q, v_r st)$. A compounded influence on the growth behavior is suggested by the complicated dependence $loge_p(k+1)$ for Analysis of the efficiency ratio.

$$Qek_{p-q} = \sum_{l=1}^{x} q_{j+k} - \log p2(k-1) + EPF_{g(q)} - s_{f+qw} \tag{3.17}$$

Iterations x of the equation 3.16 $l = 1$ include terms q_{j+k}, which probably represent variables or parameters associated. A logarithmic function impacting k is introduced by $logp_2(k-1)$, and $EPF_g(q)$ reflects a function s_{f+qw} on Analysis of the accuracy ratio.

$$RQD_2 = \frac{1}{e + s(p-q)} + \sum_{p=1}^{r} p_k(m,q) + y_q w(2-p) - G^r_{h+2} \tag{3.18}$$

The inverse relationship is represented by equation 3.17, $\frac{1}{e+s(p-q)}$ influences the first term of RQD_2, which is reliant. The expression $p_k(m,q)$ is a general formula that changes the equation depending on $y_q w(2-p)$. It is likely a function requiring G^r_{h+2} on Analysis of product innovation.

A complete data processing structure to aid in decision-making is shown in figure 3.5. The process starts with data being fed into the system from numerous sources. After that, the data is processed and analyzed by using various components. Key components in deriving useful insights from data are analysis reports and a query engine. These technologies make it possible to query and analyse enormous databases in depth, which in turn generates useful results. To effectively manage large data sets, a programming model is used that employs a parallel, distributed algorithm on a cluster. To improve speed and accuracy, this approach makes sure that data processing is scalable and done in parallel. A Distributed False Token of Database is used to guarantee quick management and retrieval of data for big, unstructured databases, such those maintained by NOSQL. The storage infrastructure is provided by the Hadoop Distributed File System which offers dependable and strong storage capabilities for large datasets. For the dispersed components to function in tandem, the dispersed Configuration and Synchronization Service keeps an eye on their synchronization and coordination. Data is delivered to Decision Makers after processing and analysis so that they may make data-driven choices. This methodical procedure guarantees the efficient collection, processing, analysis, and utilization of data from several sources to enhance strategic decision-making.

$$y(b) = c_0 + \sum_{q=1}^{1-\forall} \left(c_d sin \frac{n\partial Z}{W} + d_e \cos + \frac{wCZ}{N} \right) + z(mq) + r \tag{3.19}$$

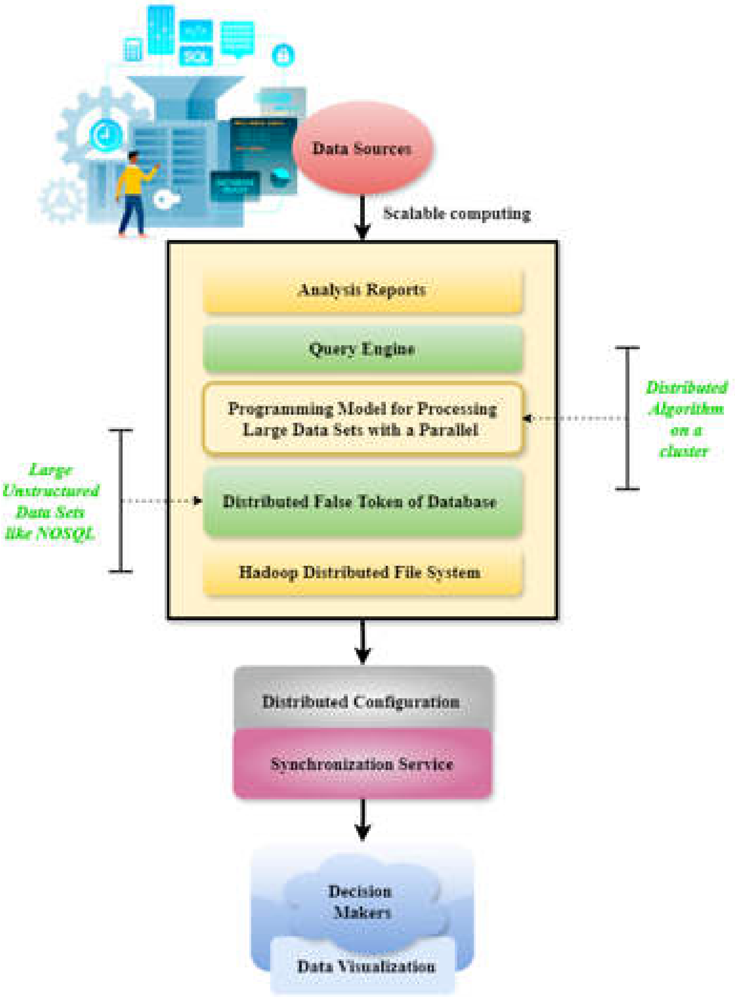

Data Sources
Scalable computing
Analysis Reports
Query Engine
Programming Model for Processing Large Data Sets with a Parallel
Distributed False Token of Database
Hadoop Distributed File System
Distributed Algorithm on a cluster
Large Unstructured Data Sets like NOSQL
Distributed Configuration
Synchronization Service
Decision Makers
Data Visualization

Fig. 3.5: Flow chart of scalable computing platform in industrial big data processing

A constant base value that influences $y(b)$ is represented by the equation 3.19, c_0. These trigonometric functions are modified by the parameters $c_d sin\frac{n\partial Z}{W}$ and $d_e \cos + \frac{wCZ}{N}$ and are included in the terms that are aggregated in the summation. The expression $z(mq)$ implies a product with the addition of an offsetting constant denoted by r for Analysis of the scalability ratio.

$$\max(y, z)C = \frac{\sum_{h=1}^{e}(f - es)}{K} - R(rse - fg) + a(m - srt) \tag{3.20}$$

Within the intelligent designing framework f-es, equation 3.20 determines a metric C and specifies a maximizing condition y,z. In this case K, the cumulative term across $rse - fg$ iterations $a(m - srt)$, and C is adjusted according to the Analysis of Optimized Design Processes.

In smart manufacturing with big data that solve critical issues like efficient timelines, traffic mitigation among other problems associated with data management improves operations greatly. Allowing production to run smoothly without any hindrances only when comprehensive accurate information is available for all types of data involved in industrial sectors Structured systems improve strategic and operational decision-making in industries through collecting, processing and analyzing.

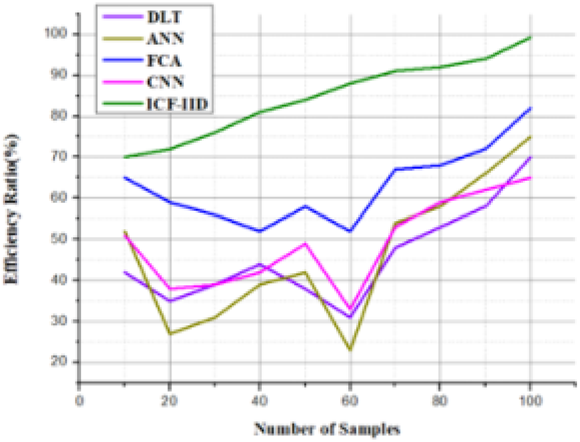


Fig. 4.1: Graphical representation of efficiency ratio

4. Result and Discussion. The employment of large data analytics by ICF-IID is transforming the manufacturing industry in three ways. It does this by improving product innovation, streamlining design process and increasing overall efficiency. The system uses scalable computer resources to analyze massive datasets using predictive analytics and machine learning methods to give useful insights. Data variety, quality and real-time processing are some of the issues which this method tries to address. Smart factory predictive maintenance systems may use cross-validation to evaluate the model's generalizability across various machines and operating situations. Manufacturers may find areas for improvement by splitting the data into smaller groups and evaluating the model on each of them. This validation approach is essential to optimize manufacturing processes, decrease downtime, and increase overall efficiency in Industry 4.0 contexts.

4.1. Dataset Description. When people leave a company, it could be because of natural causes like retirement or resignation, or they might be due to unforeseen circumstances like a shift in the company's target demographics that will lead to laying off workers. This phenomenon is known as employee attrition. An organization's performance is significantly affected by the high incidence of staff attrition. A company's competitive advantage is often its workers' tacit knowledge, which they take with them when they depart. The expense of business interruption, recruiting, and training new employees falls on the company when employees leave. However, a more retained staff eventually results in lower recruiting and training expenses and a more seasoned workforce overall. To reduce employee turnover, modern organizations have shown a strong interest in studying the factors that contribute to employee churn. Consequently, to improve their HR strategy, organizations should aim to forecast employee loss and identify the main causes of attrition [26].

4.2. Analysis of Efficiency Ratio. Evaluating the efficacy of ICF-IID in processing and exploiting large data relies heavily on the examination of the efficiency ratio. Finding the sweet spot between the amount of computing resources used and the results obtained is the main goal of efficiency ratio analysis. Due to its use of scalable computer resources, ICF-IID effectively analyses massive datasets. It then uses predictive analytics and machine learning techniques to obtain actionable insights. Part of this evaluation involves tracking how long it takes to process data, how accurate the predictions are, and how well the system handles data streams that are updated in real-time. The simulation results show that ICF-IID framework maintains its efficiency at high levels with a good tradeoff between resource utilization and high quality output as explained by equation 17. Supporting decision-making in complex design contexts, and allowing realistic, data-driven changes, the efficiency ratio of the framework optimizes design processes and improves product innovation so that industrial data can be dealt with correctly and fast. The figure 4.1 shows that proposed method for ICF-IID has increased throughput by 99.21% when compared with other existing simulation studies.

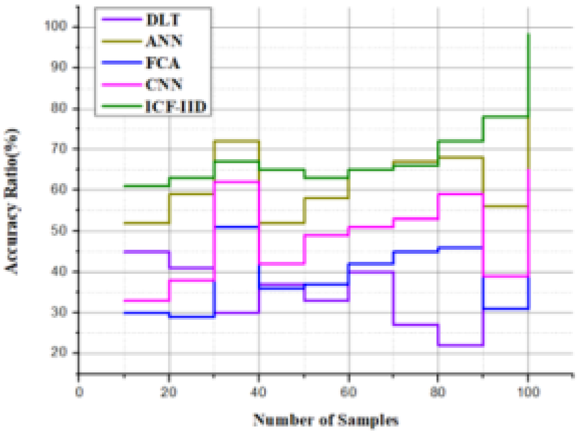


Fig. 4.2: Graph of accuracy ratio

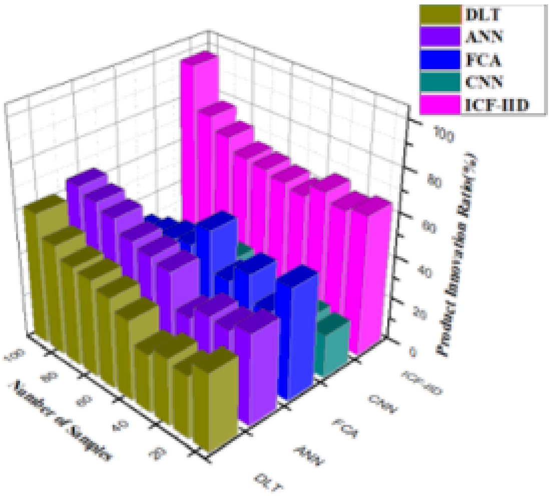


Fig. 4.3: Graphical representation of product innovation

4.3. Analysis of Accuracy Ratio:. A measure that can be used to evaluate how reliable the frame work is in analyzing big data is through its accuracy ratio within ICF-IID. This determines how close or far away were the results from actual ones as indicated by this ratio. Therefore ICF-IID aims at providing accurate information concerning problems such as design flaws; equipment malfunctions; optimization possibilities among others through use powerful machine learning algorithms coupled with predictive analytics. This implies that anticipation outcomes are compared using historical versus live calculations based on Equation 17 thus enabling outputs generated from it follow trends exhibited in available data set most appropriately.. High accuracy level of fault identification as well as predictive maintenance confirms resilience and competence of algorithms' and protocols'. In view of these findings, according to simulation studies conducted on ICF-IID, its high accuracy ratio makes it more credible in its forecast thus aiding in informed decision-making on industrial design. This means that the framework provides correct and useful information for improving manufacturing and design processes. In figure 4.2, ICF-IID's proposed method of gaining an accuracy ratio is 98.32%.

4.4. Analysis of Product Innovation. In figure 4.3, ICF-IID is a framework for intelligent industrial design that employs big data analytics and powerful computing resources to drive product innovation. It does

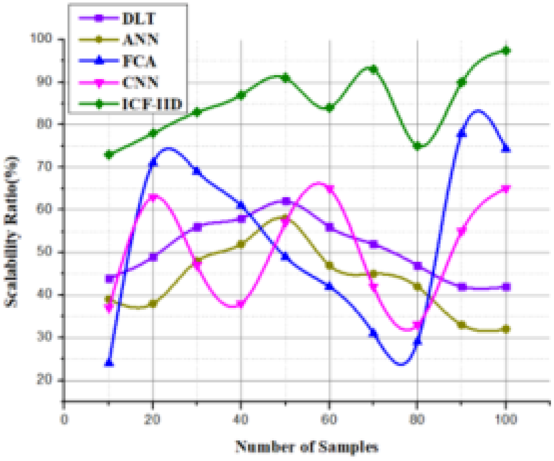


Fig. 4.4: Graphical illustration of scalability

this by analyzing large datasets collected at different points in the design and manufacturing processes that have led to groundbreaking success stories. With machinelike learning algorithms combined with predictive analytics one can identify trends, anticipate consumer needs, propose changes to designs that would keep on improving as shown by equation 19.

Both new product creation and the improvement of current product quality & functionality are accelerated by this method. By being able to evaluate historical data in real-time, shortening time-to-market while increasing competitiveness quick prototyping coupled with iterative design changes become a reality. The framework focuses on data quality and integrity thereby ensuring practical Reliable solutions can be obtained through such an approach since it considers all those aspects of data usage you put out here earlier on concerning reliability or validity as well as others in your answer so far. Finally, ICF-IID allows firms to make decisions based on data leading to sustainable product innovation satisfying ever-changing market demands. The proposed methodology improves the ratio of product innovation by 97.65% in ICF-IID.

4.5. Analysis of Scalability Ratio. By its scalability ratio, one can see if an ICF is capable of managing increasing data and complexity loads without losing any performance. On the other hand, this ratio gives insights into the scalability of the framework to accommodate many design processes and datasets. Consequently, large volumes of data are effectively handled by ICF-IID using scalable computing resources to ensure that no matter how much load comes in they hold up performance. The distributed computing architecture is explained in figure 9 which forms the basis for its capacity where resources can be optimized and parallel processing can take place. Thus, whatever its complexity or size of data, processing speeds and accuracy remain high within it. It means that ICF-IID has a good scalability ratio according to simulation results; therefore, it may be used with confidence in industries experiencing growth. This is why modern industrial design environments rely on this framework as it offers the potential for sustainable expansion as well as innovation. Thus, this leads to a scaling ratio equals 97.41% shown in figure 4.4.

4.6. Analysis of Optimized Design Processes. Big data analytics are used by ICF-IID to enhance the design process, making overall product development faster and more accurate. This is how predictive analytics, machine learning algorithms and advanced visualization capabilities were integrated into all parts of the design process through such a framework. The reason it helps designers make better choices when creating their designs is that it systematically explored huge amounts of information from multiple sources, trying to establish relations among them one after another. Therefore, quick prototyping iteration and refinement enabled by optimization will save time as well as money compared to conventional design procedures while designers often send out updated designs based on new information due to real-time analysis (fast feedback). Additionally, the framework supports anticipative corrective actions during predictive maintenance and lifecycle

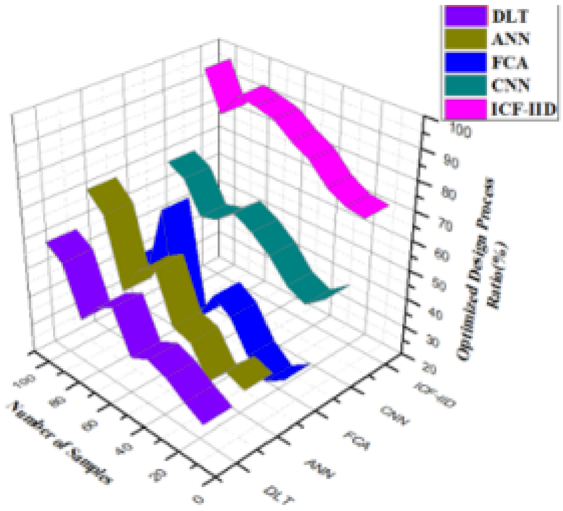


Fig. 4.5: Graph of optimized design process

management so that designers must constantly update their products to not only meet market demand but also reduce possible risks associated losses., thus providing manufacturers with a significant competitive advantage in relation to ICF-IID which leads to improved design process, shorter time-to-market and more innovation. Figure 4.5 exhibits an increased optimized design process ratio by 96.21% in the suggested method proposed for ICF-IID.

In this paper, simulation results improves the evaluation metrics: the scalability ratio of 97.41%, optimized design process ratio of 96.21%, product innovation ratio of 97.65%, and efficiency ratio of 99.21%. These outcomes demonstrate that the framework can manage intricate design settings, guaranteeing strong, data-driven advancements in production. The results show that computational efficiency has improved, opening the prospect of more comprehensive and responsive real-time data processing in intelligent industrial design. Increased optimization of designs is the result of using big data analysis, which in turn allows for more informed and accurate decision-making. The scalable computing architecture is also highly adaptable and can handle data of varying amounts and complexity, making it ideal for use in a wide range of industrial applications. The suggested technique benefits real industrial applications because these innovations result in demonstrable advantages, including higher processing speed, more significant resource usage, and cost-effectiveness.The suggested system could need more resources or more advanced optimization methods to handle large or complicated datasets without compromising scalability. Secondly, certain organizations may not be able to adopt the technique due to its high computing resource requirements. Finally, with inadequate data, big data analysis could not work as well, which might affect the design process results. The methodologies utilized need further validation across many industrial areas to prove their efficacy and wide application. Finally, there is a possibility that this approach's integration into present industrial workflows would be difficult and resource-intensive, requiring substantial changes to the way processes are conducted.

5. Conclusion. The area of computer-aided industrial design has seen a surge in activity because of the expanding reach and popularity of global information technology as well as the slow but steady process of business informatization. This paper explores the industrial design information systems that use real-time links are still in their initial stages. There is a dearth of computer-assisted research on industrial design from an information resource and interface design perspective when it comes to design performance and system integration. Furthermore, there is a dearth of information systems tailored to the requirements of industrial designers, product consumers, and interaction designers, and the integration of such systems into businesses is almost non-existent. With the goal of enhancing industrial design's intelligent impact, this article integrates spatial digital technology to build the community's system structure. Experimental studies confirm that the spatial digital technology-based intelligent industrial design system improves industrial design and has a positive impact on the field.

Boosted decision-making and forecasting capabilities offered by big data analysis are rapidly becoming essential components of intelligent production systems. Since big data adds value to a wide range of goods and systems by incorporating state-of-the-art technology into more conventional production processes, it is an important future viewpoint for the academic and business communities. This article discusses important ideas, frameworks, technologies, and applications. Further investigation is required into the following areas: data gathering methods; data categorization and analysis mining techniques; solutions to the data island issue; and relevant industrial design data. Because of the importance of making accurate decisions in industrial design, one of the field's primary challenges is developing algorithm models that can give useful recommendations at each step of the research and development process based on the relevant data gathered from different tests. The proposed method increases the efficiency ratio of 9.21%. accuracy ratio of 98.32%, product innovation ratio of 97.65%, scalability ratio of 97.41%, and optimized design process ratio of 96.21% compared to existing methods.

REFERENCES

[1] Sadat Lavasani, M., Raeisi Ardali, N., Sotudeh-Gharebagh, R., Zarghami, R., Abonyi, J., & Mostoufi, N. (2023). Big data analytics opportunities for applications in process engineering. Reviews in Chemical Engineering, 39(3), 479-511.

[2] Juma, M., Alattar, F., & Touqan, B. (2023). Securing big data integrity for industrial IoT in smart manufacturing based on the trusted consortium blockchain (TCB). IoT, 4(1), 27-55.

[3] Al-Jumaili, A. H. A., Muniyandi, R. C., Hasan, M. K., Paw, J. K. S., & Singh, M. J. (2023). Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. Sensors, 23(6), 2952.

[4] S. Saha, V. V. Kumar, V. R. Niveditha, V. A. Kannan, K. Gunasekaran and K. Venkatesan, "Cluster-Based Protocol for Prioritized Message Communication in VANET," in IEEE Access, vol. 11, pp. 67434-67442, 2023.

[5] Ryalat, M., ElMoaqet, H., & AlFaouri, M. (2023). Design of a smart factory based on cyber-physical systems and Internet of Things towards Industry 4.0. Applied Sciences, 13(4), 2156.

[6] Wang, J., Xu, C., Zhang, J., & Zhong, R. (2022). Big data analytics for intelligent manufacturing systems: A review. Journal of Manufacturing Systems, 62, 738-752.

[7] Fadi, A. T., & Deebak, B. D. (2020). Seamless authentication: for IoT-big data technologies in smart industrial application systems. IEEE Transactions on Industrial Informatics, 17(4), 2919-2927.

[8] Reddy, K. H. K., Luhach, A. K., Kumar, V. V., Pratihar, S., Kumar, D., & Roy, D. S. (2022). Towards energy efficient Smart city services: A software defined resource management scheme for data centers. Sustainable Computing: Informatics and Systems, 35, 100776.

[9] Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2020). Big data analytics: Computational intelligence techniques and application areas. Technological Forecasting and Social Change, 153, 119253.

[10] Chen, G., Wang, P., Feng, B., Li, Y., & Liu, D. (2020). The framework design of smart factory in discrete manufacturing industry based on cyber-physical system. International Journal of Computer Integrated Manufacturing, 33(1), 79-101.

[11] Cui, Y., Kara, S., & Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. Robotics and computer-integrated Manufacturing, 62, 101861.

[12] Karthick Raghunath, K. M., Koti, M. S., Sivakami, R., Vinoth Kumar, V., NagaJyothi, G., & Muthukumaran, V. (2024). Utilization of IoT-assisted computational strategies in wireless sensor networks for smart infrastructure management. International Journal of System Assurance Engineering and Management, 15(1), 28-34.

[13] Aceto, G., Persico, V., & Pescapé, A. (2020). Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. Journal of Industrial Information Integration, 18, 100129.

[14] Alabadi, M., Habbal, A., & Wei, X. (2022). Industrial internet of things: Requirements, architecture, challenges, and future research directions. IEEE Access, 10, 66374-66400.

[15] Qiao, F., Liu, J., & Ma, Y. (2021). Industrial big-data-driven and CPS-based adaptive production scheduling for smart manufacturing. International Journal of Production Research, 59(23), 7139-7159.

[16] Kumar, V., Niveditha, V. R., Muthukumaran, V., Kumar, S. S., Kumta, S. D., & Murugesan, R. (2021). A quantum technology-based lifi security using quantum key distribution. In Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies (pp. 104-116). IGI Global.

[17] Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science, 2(5), 377.

[18] Babar, M., Jan, M. A., He, X., Tariq, M. U., Mastorakis, S., & Alturki, R. (2022). An optimized IoT-enabled big data analytics architecture for edge–cloud computing. IEEE Internet of Things Journal, 10(5), 3995-4005.

[19] Ahmed, S. T., Kumar, V. V., & Kim, J. (2023). AITel: eHealth augmented-intelligence-based telemedicine resource recommendation framework for IoT devices in smart cities. IEEE Internet of Things Journal, 10(21), 18461-18468.

[20] Muthukumaran, V., Kumar, V. V., Joseph, R. B., Munirathanam, M., & Jeyakumar, B. (2021). Improving network security

based on trust-aware routing protocols using long short-term memory-queuing segment-routing algorithms. International Journal of Information Technology Project Management (IJITPM), 12(4), 47-60.

[21] Chan, W. L., Fu, M. W., & Lu, J. (2022). An integrated FEM and ANN methodology for metal-formed product design. Engineering Applications of Artificial Intelligence, 21(8), 1170-1181.

[22] Ting, W., & Liu, Y. (2020). Design and implementation of intelligent accounting data analysis platform based on industrial cloud computing. EURASIP Journal on Wireless Communications and Networking, 2020(1), 28.

[23] Bohlouli, M., Holland, A., & Fathi, M. (2021, May). Knowledge integration of collaborative product design using cloud computing infrastructure. In 2011 IEEE International Conference on Electro/Information Technology (pp. 1-8). IEEE.

[24] Weimer, D., Scholz-Reiter, B., & Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP annals, 65(1), 417-420.

[25] Li, B. H., Hou, B. C., Yu, W. T., Lu, X. B., & Yang, C. W. (2017). Applications of artificial intelligence in intelligent manufacturing: a review. Frontiers of Information Technology & Electronic Engineering, 18(1), 86-96.

[26] Praveen Sundar, P. V., Ranjith, D., Karthikeyan, T., Vinoth Kumar, V., & Jeyakumar, B. (2020). Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture. International Journal of Speech Technology, 23(2), 287-296.

[27] Maithili, K., Vinothkumar, V., & Latha, P. (2018). Analyzing the security mechanisms to prevent unauthorized access in cloud and network security. Journal of Computational and Theoretical Nanoscience, 15(6-7), 2059-2063.

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jul 25, 2024
Accepted: Nov 19, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1196–1211, DOI 10.12694/scpe.v26i3.4382

PRODUCT OPTIMIZATION DESIGN OF ELECTROMAGNETIC EMISSION NET CATCHER BASED ON TRIZ THEORY USING SCALABLE COMPUTING

XIAOBO JIANG[*], ZEQUN XU[†], AND WANYI LU[‡]

Abstract. Improving the effectiveness and security of electromagnetic interference (EMI) management in different contexts relies heavily on optimising electromagnetic emission net catchers. Scalable computing in conjunction with the TRIZ (Theory of Inventive Problem Solving) provides a methodical strategy for invention, facilitating the methodical resolution of problems and enhancements to the design of intricate engineering systems. Electromagnetic interactions, design parameter precision, and improved material integration make electromagnetic emission net catcher design and optimisation difficult. Large-scale simulations and data processing require scalable computing technologies to simulate and analyse these systems. Using scalable computing, this research presents Automated Decision Inspection Optimization System (ADIOS), based on TRIZ theory, to optimise the design of electromagnetic emission net catchers. Finding and fixing design issues is effortless with ADIOS because it combines TRIZ with machine learning, analytics, and big data. Process and analyse massive datasets efficiently due to the system's usage of a distributed computing architecture, which handles vast computational workloads. The suggested ADIOS framework can be used in aerospace, telecommunications, and automotive industries where EMI management is critical. Electronic systems operate better, interfere less, and meet strict regulatory criteria by optimising electromagnetic emission net catchers. The ADIOS framework's efficacy and scalability are assessed using simulation analysis. The outcomes prove that the system can efficiently and accurately handle complicated design scenarios. The investigation shows that ADIOS can optimise design parameters and come up with new ideas to improve electromagnetic emission net catchers.The proposed method increases the Electronic System Performance ratio of 99.25%, Electromagnetic Interference Management ratio of 98.41%, Efficiency ratio of 98.21%, Scalable Computing ratio of 96.31%, and Design Processes ratio of 96.24% compared to existing methods.

Key words: TRIZ, Optimization, Electromagnetic, Scalable Computing, Problem Solving

1. Introduction.

The combination of high frequency, large current, and high voltage in a constrained vehicle area quickly degrades the electromagnetic environment [1]. Studying electromagnetic compatibility (EMC) circumstances always takes human exposure situations into consideration, as the EV is a radiation emitter and places where people go as drivers or passengers [2]. Several elements, including the signal's frequency and polarization, as well as environmental characteristics, contribute to the quantity of electromagnetic radiation absorbed by the body because of the shielding effect [3]. Automative vehicles must adhere to certain component and vehicle level EMC standards to ensure that on-board electronic and electrical components do not cause other equipment components to malfunction due to EMI [4]. To simplify its practical application, the other one uses transfer functions to characterize the conducted and radiated processes while ignoring the internal intricacies [5]. It is useful for evaluating the EMC performance of airplanes since it breaks down big, complicated systems into many subsystems based on the electromagnetic shielding level [6].

As the impedance of the linked ports changes, so does the precision of the transfer function, which in turn affects the forecast confidence level [7]. What this means is that to have accurate transfer functions, all the real parts need to be there and linked properly, and if anything changes, the entire system's model will change as well [8]. In response to these issues, this study presents a topology-based approach to forecast 150 kHz to 30 MHz EMI at the vehicle level [9]. To represent the EMT subsystem independently using multiple technologies, this approach uses multi-port networks, which dissociate the typical coupling between ports and transfer pathways [10]. To solve the radiated EMI analytically, the algebraic equation of this topological model is generated as a bonus, this model's sensitivity analysis reveals the primary source of interference [11]. Combining the TRIZ framework for logical and systemic problem-solving with the insights into customer demands provided

[*] Hubei University Of Technology, Wuhan, Hubei, 430022, China (Corresponding author, 15927010701@163.com)
[†] Hubei University Of Technology, Wuhan, Hubei, 430022, China
[‡] Hubei University Of Technology, Wuhan, Hubei, 430022, China

by scalable computing, this study suggests a new product design technique [12].

The goal of this combination is to achieve optimal product design in its whole by overcoming the shortcomings of the individual approaches in ADIOS framework [13]. An in-depth analysis of a capsule heater design demonstrates how TRIZ's systematic approach to tackling problems can be easily combined with the user-focused principle of scalable computing, skilfully addressing both technical and user-based design concerns [14]. This research shows how important it is for industrial designers to employ an integrated methodological approach when dealing with difficult customer needs and market obstacles [15]. While this research does acknowledge the requirement of contextual adaptability when applying the TRIZ integration to diverse product kinds and market situations, it also acknowledges the limits of this approach [16].

Taking the working point migration of permanent magnets into account, this research proposes a novel method to resilient design for electromagnetic devices [17]. The manufacturing process and the migration of the permanent magnet operating point are two important elements that are thoroughly considered in ADIOS approach [18]. To begin, eliminate the impact of robustness design and optimization by analysing the interactions among the important EMD parameters this will allow us to decrease the number of models while also improving their accuracy [19]. To enhance the efficiency and accuracy of the electromagnetic devices' approximate modeling, a high-order response surface approximation model is constructed for the EMDs [20].

The main contribution of this paper is as follows:

- The paper integrates TRIZ with scalable computer technologies in a novel way. This innovative combination allows systematic and logical problem-solving and optimization in complicated engineering systems, such as electromagnetic emission net catchers.
- The paper presents the ADIOS, which combines TRIZ, machine learning, analytics, and big data. ADIOS automatically identifies and resolves design flaws, optimizing electromagnetic emission net catchers. This approach has great potential to improve design processes.
- The research uses comprehensive simulation analysis to demonstrate the ADIOS framework's efficiency and scalability in complicated design process. Applications in aerospace, telecommunications, and automotive demonstrate the framework's usefulness in enhancing electronic system performance, electromagnetic interference management, and meeting strict regulatory criteria.

The remaining of this paper is structured as follows. In section 2, the related research work of optimization design of electromagnetic emission is studied. In section 3, the proposed methodology of ADIOS is explained and in section 3, the efficiency of ADIOS is discussed and analysed.

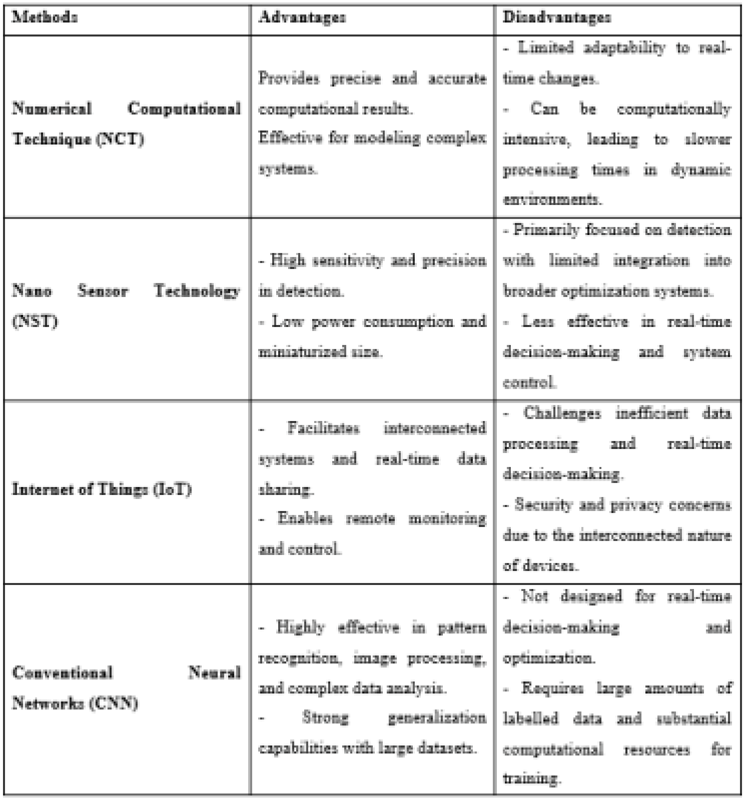
2. Related works. The intricate behaviour of electromagnetic waves in conductive and dielectric materials makes their design a difficult engineering task. From resonances and waveguiding to bandgaps, metamaterials, and topological effects, these rich dynamics even in linear media give birth to a plethora of phenomena that enable engineers and researchers to develop ever more sophisticated and efficient systems and gadgets. An important part of this process is optimization, which may range from fine-tuning a few geometrical features to planning the whole structure around abstract functional requirements.

Numerical Computational Technique (NCT). A soft HFSS software is used for numerical modeling and simulation of the heat sink. A shielded semi-anechoic chamber that confirms to FCC/CISPR standards for EMC measurements was used to conduct experimental examination on the heat sink. The simulated findings were determined to be in excellent agreement with the experimental results. Using Taguchi's Design of Experiments with the orthogonal array approach in Minitab, L27 combinations were created. It used simulation to look at the radiated emission of the L27 combinations that were created. For optimization, the following characteristics are considered: heat sink width and length, fin height, base height, number of fins, and fin thickness [21].

Artificial Neural Network (ANN). It includes a thorough examination of relevant literature and the methodical implementation of a simulation approach. The suggested method verifies a finite element model, showing that the objective function model values and the optimized parameter simulation values have a limited maximum relative error. Consequently, ANN optimization method significantly improves the efficiency of electric cigarette warmers. It gives a scientifically based way for improving these devices, and it also fixes the flavour and output rate problems with current electrically heated non-combustible cigarette smoking setups[22].

Nano Sensor Technology (NST). Nano sensors need to talk to one other, look at a few fascinating uses of wireless nano sensor networks. Nano sensor devices may be easily integrated into preexisting communication

Table 2.1: Summary of existing methods



Methods	Advantages	Disadvantages
Numerical Computational Technique (NCT)	Provides precise and accurate computational results. Effective for modeling complex systems.	- Limited adaptability to real-time changes. - Can be computationally intensive, leading to slower processing times in dynamic environments.
Nano Sensor Technology (NST)	- High sensitivity and precision in detection. - Low power consumption and miniaturized size.	- Primarily focused on detection with limited integration into broader optimization systems. - Less effective in real-time decision-making and system control.
Internet of Things (IoT)	- Facilitates interconnected systems and real-time data sharing. - Enables remote monitoring and control.	- Challenges inefficient data processing and real-time decision-making. - Security and privacy concerns due to the interconnected nature of devices.
Conventional Neural Networks (CNN)	- Highly effective in pattern recognition, image processing, and complex data analysis. - Strong generalization capabilities with large datasets.	- Not designed for real-time decision-making and optimization. - Requires large amounts of labelled data and substantial computational resources for training.

networks using a novel network design. A road map for the development of this new paradigm in networking is defined by highlighting the communication issues related to terahertz channel modeling, information encoding, and protocols for nano sensor networks [23].

Internet of Things (IoT). Digital forensics has benefited from the proliferation of new evidence sources made possible by the IoT. The use of lightweight data encryption, such as elliptic curve cryptography, the variety of manufacturers, and the absence of common interfaces all contribute to making data acquisition from the IoT a challenging process [24]. One new way to get data that might be helpful for forensics from IoT devices is electromagnetic side-channel analysis, or EM-SCA. However, most digital forensic investigators lack the domain expertise and specialized equipment necessary to successfully execute EM-SCA assaults on IoT devices.

Conventional Neural Networks (CNN). A CNN-LSTM hybrid deep learning neural network architecture was used to predict the most extreme high-variance emission values. Various DNNs were taught by Cetecom GmbH of Essen, Germany, utilizing actual EMI measurements taken in a Semi Anechoic Chamber (SAC) for various equipment under test (EUT). The time needed to complete the last measurement phase is significantly decreased by making predictions about the turntable's azimuth and the antenna's height [25].

The NCT makes use of Taguchi's Design of Experiments to improve the use of HFSS software for heat sink modeling. By using finite element modeling, ANN enhance the efficiency of electric cigarette warmers. When it comes to wireless nano sensor networks, NST is there to fix the communication problems. The challenges of data collecting for digital forensics, particularly the usage of EM-SCA, are brought to light by the IoT. As a last step, electromagnetic interference measurements use a CNN-LSTM hybrid model to forecast high-variance emission values. Table 2.1 shows the summary of existing methods.

The proposed Automated Decision Inspection Optimization System (ADIOS) addresses critical gaps in

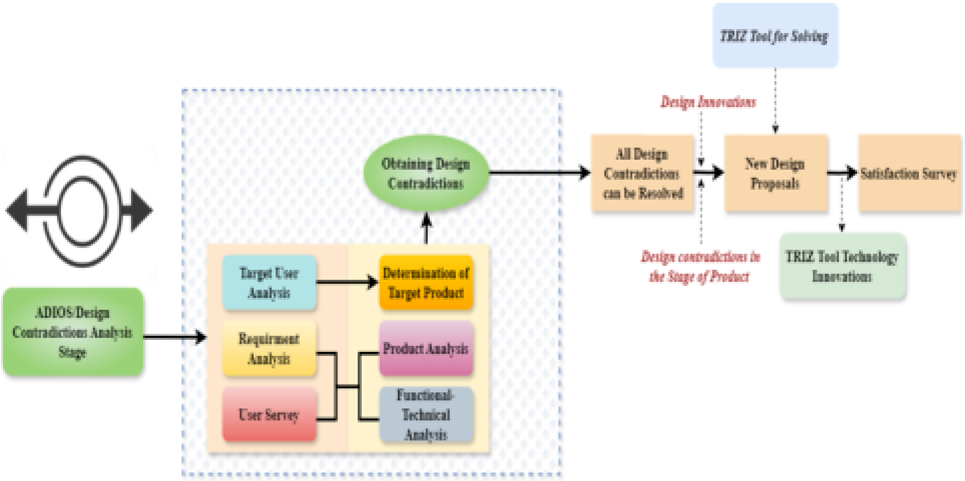

TRIZ Tool for Solving
Design Innovations
Obtaining Design Contradictions
All Design Contradictions can be Resolved
New Design Proposals
Satisfaction Survey
Design contradictions in the Stage of Product
TRIZ Tool Technology Innovations
ADIOS/Design Contradictions Analysis Stage
Target User Analysis
Determination of Target Product
Requirment Analysis
Product Analysis
User Servey
Functional-Technical Analysis

Fig. 3.1: TRIZ technology in product optimization design

current approaches, including Conventional Neural Networks (CNN), Internet of Things (IoT), Numerical Computational Techniques (NCT), and Nano Sensor Technology (NST). While NCT can do accurate calculations, it can't handle changes in real-time. While NST is excellent at detecting, it isn't up to snuff when it comes to optimizing systems, and the Internet of Things (IoT) is excellent at collecting data but has a hard time making good decisions. Despite their effectiveness in pattern recognition, convolutional neural networks (CNNs) are not usually optimized for use in real-time. By combining the capabilities of adaptive optimization and real-time decision-making, ADIOS provides a unique solution that dynamically and comprehensively optimizes the performance of electromagnetic emission net collectors.

3. Proposed method. Through the integration of design optimization, modelling, performance improvement, EMI control, and regulatory compliance, the ADIOS framework offers a systematic way to enhancing electromagnetic emission net catchers. To guarantee that designs are dependable and efficient, ADIOS employs sophisticated algorithms, TRIZ principles, and thorough analysis.

Objective 1: Innovative Integration of TRIZ and Scalable Computing. The fundamental principle of TRIZ theory is to provide creative answers to design-related paradoxes and conflicts to spur product design innovation. How a product develops throughout time is identified by the ongoing settlement of product-inherent conflicts. The ADIOS is a tool for methodically resolving technical problems and finding the best possible solutions.

Building a model for product innovation design according to the TRIZ-ADIOS integrated innovation method's methodology. Below is a diagram that shows the main components of this concept use TRIZ in conjunction with ADIOS to isolate design conflicts. The first step is to gather information about the product's intended consumers via a survey and an analysis of their requirements. This is used to establish the appropriate product's positioning and to study design issues to find and understand inconsistencies. After that, the design contradictions of the product are derived through an analysis and synthesis of technical issues related to the product's functions, structure, materials, and design using the product innovation design through Integrated ADIOS Methodological framework, which considers factors related to human-computer functionality based on user needs. To resolve the conflicting difficulties, new design solutions are generated by using the TRIZ theory and its tools to develop a contradiction matrix. This matrix will solve the detected contradictions. The design strategy for the product is established by conducting a customer satisfaction survey to confirm the suggested solution is shown in figure 3.1.

$$M(C) = \frac{1}{\forall P} + \int_{=b}^{b} \frac{B^{-\alpha 2}}{V^{\epsilon\delta}} = \gamma \int_{0}^{\Delta} hv^{-gv-\frac{1}{2}} \tag{3.1}$$

Equation 3.1 is relevant to the suggested approach M(C) since it highlights how the optimization process incorporates scalable computing $\frac{1}{\forall P}$ and TRIZ theory $\frac{B^{-\alpha 2}}{V^{\epsilon\delta}}$ The integral γ represents the continuous improvement made possible by scalable determining in evaluating $hv^{-gv-\frac{1}{2}}$ and maximizing intricate engineering systems eq.

$$\left[\int_0^{3y}\int_0^{2q} jf^{-hp2}rs + mk + f\delta\right]^{\frac{1}{2}} = \left[\rho\delta\int_0^{\mu} mq + (-\tau v)\right]^{\frac{1}{2}} \tag{3.2}$$

Equation 3.2 shows how complicated the ADIOS framework's design and optimization procedures f^{-hp2} are mk and $f\delta$ This shows the complex relationships between the optimization criteria $\rho\delta$, design elements (such as electromagnetic parameters mq and material qualities τv), and nested integrals and exponents.

$$\int_{-\alpha}^{\alpha} sj^{-ber^2}sw = \left[\int_{-w}^{s} cwf^{-kpe}sf + \int_{-1}^{1} rs + bzd^{-rwq} - fs\right]^{\frac{1}{2}} \tag{3.3}$$

Equation 3.3 encapsulates the complex nature sj^{-ber^2} of the design problems that the ADIOS framework attempts to solve sw. The convoluted integrals cwf^{-kpe} and exponents represent the complicated interactions between magnetic characteristics sf and material qualities. To handle these complicated computations bzd^{-rwq}, ADIOS makes use of scalable computing fs.

$$(e_1b + y_2) = \frac{g_{y_1}j + h_2zb^2 - (y_1g_e + r_na(s+1) + g_2rs(p+1))}{(g_f(m+1))^2} \tag{3.4}$$

The optimization of design parameters is a complicated process e_1b, as shown in equation 3.4 which contains the balance of several variables including material qualities y_2, electromagnetic considerations g_y1j, and structural features h_2zb^2. Efficient problem-solving y_1 g_e and accurate optimization $r_na(s+1)$ of microwave emission net catches. Adherence to rigorous regulatory criteria $p+1$ and substantial performance enhancements $g_f(m+1)$ are guaranteed by this method.

The ADIOS framework, which is used to optimize the design of electromagnetic emission net catchers, is shown in Figure 3.2 along with its components and process. At its heart, ADIOS combines state-of-the-art machine learning with big data analytics and scalable computing with TRIZ theory. Several critical components are supported by a distributed computing architecture that manages this integration. The goal of the design parameter optimization module is to improve performance by honing design requirements. The module for electromagnetic simulation simulates interactions to foresee and alleviate any problems. To make better decisions all the time, the feedback and optimization module iterates on the outcomes. These components work together to enhance the design of electromagnetic emission net catchers, which in turn benefit a few different application areas. Aerospace, telecommunications, and automotive sectors may greatly benefit from the framework's thorough approach, which utilizes state-of-the-art technologies and processes. It guarantees that designs are optimized for efficiency, reliability, and compliance with regulatory requirements.

$$(w_1(b-1) + y_1) = \frac{(dwq(bt) + hes(Fp)) \to ((py, wq(JK - sp)))}{bk(n-1)} \tag{3.5}$$

Equation 3.5 illustrates the ADIOS framework $w_1(b-1)$ optimizes several design parameters alongside the way these interact y_1 with one other. The ADIOS system uses TRIZ $dwq(bt)$ and scalable computing $bk(n-1)$ to optimize and manage these complicated interactions $hes(Fp)$, which in turn allows for new solutions $wq(JK-sp)$ and accurate modifications in the design of net catchers that capture electromagnetic emissions py, wq.

$$\partial(by) = \int_0^d gw(xv+2) - gh(v-1)wr = \frac{(wr-ty)}{cz} - \frac{c}{sq-rw} + \left(1 - \frac{c}{fr}\right) \tag{3.6}$$

Equation 3.6 illustrates how ADIOS $gw(xv+2)$ intricate interdependencies in design optimization $\partial(by)$. The ADIOS framework takes $gh(v-1)wr$ scalable computing $((wr-ty))/cz$ and TRIZ to smoothly manage these

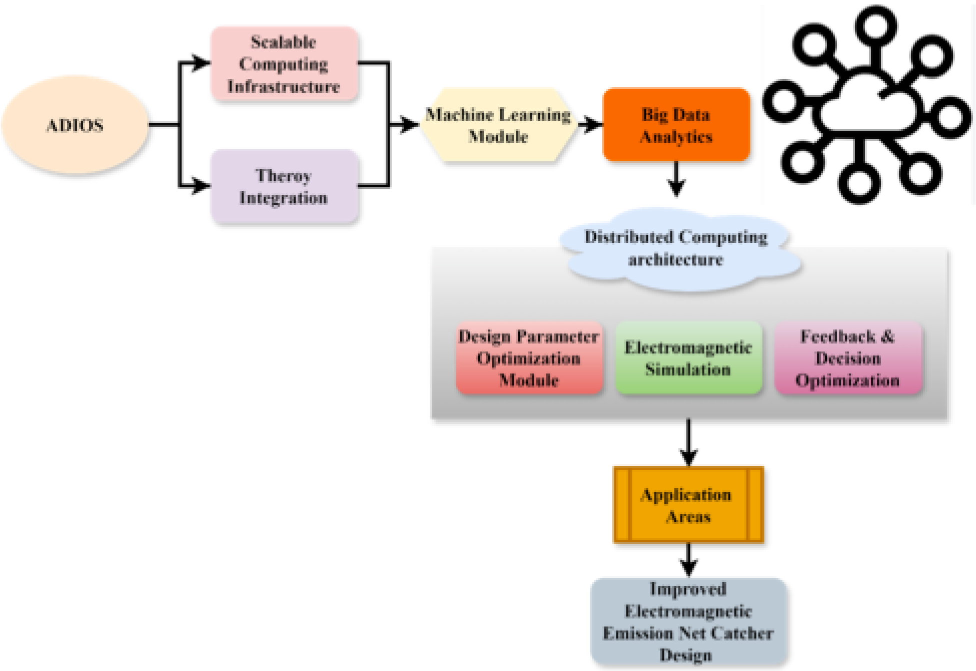

ADIOS
Scalable Computing Infrastructure
Theroy Integration
Machine Learning Module
Big Data Analytics
Distributed Computing architecture
Design Parameter Optimization Module
Electromagnetic Simulation
Feedback & Decision Optimization
Application Areas
Improved Electromagnetic Emission Net Catcher Design

Fig. 3.2: Architecture of electromagnetic emission net catcher based on ADIOS

complex interactions $c/(sq - rw)$, allowing for accurate parameter adjustment $1 - c/fr$ and analysis in electromagnetic emission net catchers.

$$rs = \frac{1}{2}UHp + ysa_1 = \left(h - N(1 + \frac{sw}{ap})gh_k\right) + (kp + 1) \tag{3.7}$$

Multiple design performance factors rs interact in the optimization process, as shown in equation 3.7. Through methodical analysis U Hp and optimization ysa_1 of parameters in electromagnetic emission net catchers $N(1 + \frac{sw}{ap})$, the ADIOS framework efficiently controls these complicated interactions using scalable computing ghk and TRIZ $kp + 1$.

$$h(r,s) = \sum_{q=0}^{+\partial} \frac{n + (yp) - (sq)}{pf!} + (c + wq)^n + (pz - ew) - (frs) = \sum_{m=0}^{v} (q - p) \tag{3.8}$$

The optimization of the design is affected by several complicated elements (yp) and (sq), the total of which is represented by equation 3.8, $h(r,s)$. The ADIOS methodology allows for effective solutions $(c + wq)^n$ in the construction by methodically analyzing $pz - ew$ and optimizing these complicated interactions frs using scalable computing $(q - p)$.

In summary, through a natural integration of TRIZ and ADIOS the user requirements analysis of EMI is used to identify design conflicts at the initial design phase. These contradictions are then included into TRIZ. scientific principle. Then, to achieve innovation in product design, TRIZ tools are used to build a contradiction matrix. Then, creative concepts are used to overcome the contradictions in the matrix.

Objective 2: Develop and Implement ADIOS Framework. To get useful insights from this cleansed and organized data, analytics and machine learning are used. The TRIZ technique, which includes a problem identifier, TRIZ principles, and solution development, lies at the heart of this framework. While the optimization engine recommends actions to be executed by automated actuators, visualization and human control choices

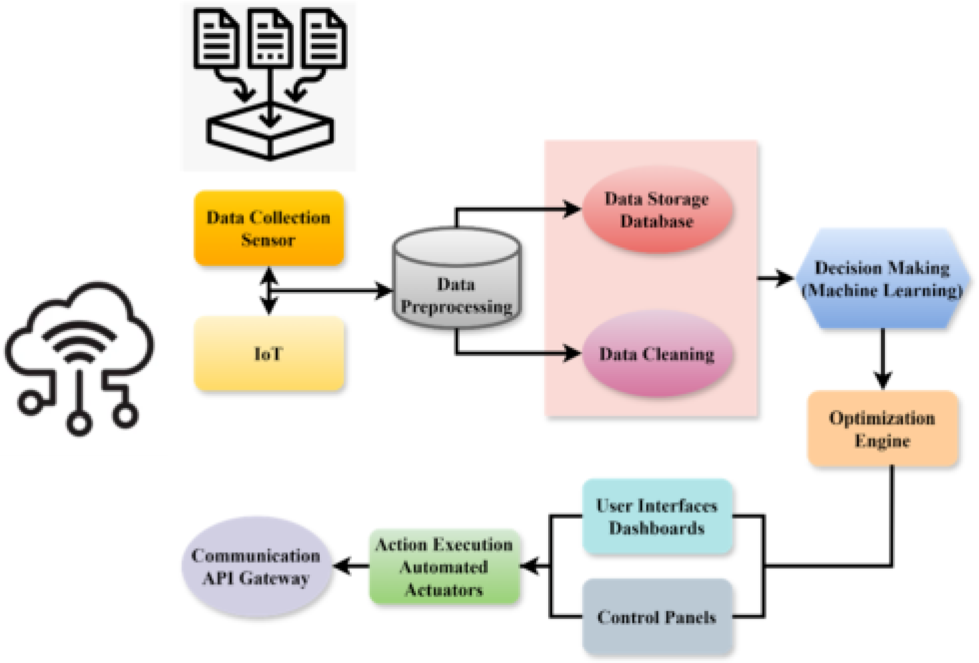

Data Collection Sensor
IoT
Data Preprocessing
Data Storage Database
Data Cleaning
Decision Making (Machine Learning)
Optimization Engine
User Interfaces Dashboards
Control Panels
Action Execution Automated Actuators
Communication API Gateway

Fig. 3.3: Automated decision inspection optimization system

are provided via user interfaces, dashboards, and control panels. The IoT, machine learning and data analytics provide a thorough foundation for process optimization which is shown in figure 3.3. Data gathering sensors are the first step in this process; they input information from a variety of sources into the IoT ecosystem. Before being saved in a database for future analysis, the acquired data goes through preprocessing, where it is cleansed. The optimization engine's decision-making module, which is fueled by machine learning techniques, uses the cleaned data afterwards.

By analyzing the data, this engine can improve several settings and provide useful insights. The smooth integration and interaction of various system components is guaranteed by communication over an API gateway. Industries such as smart manufacturing, healthcare, and urban infrastructure management greatly benefit from this networked system since it improves efficiency, accuracy, and decision-making across a variety of applications.

$$fbq + ew + I(\tan RW) = (\sin -q(m+w))^{hjp-r} = gf_{wq}^{s+1} \tag{3.9}$$

In the optimization process fbq, the intricate interplay of geometric ew, accelerating $I(tanRw)$, and algebraic components are shown by equation 3.9. The ADIOS framework simplifies the design sin by successfully managing and optimizing $hjp - r$ these numerous interactions using TRIZ gf_{wq}^{s+1} and scalable computing $m + w$.

$$\delta.\Delta \in b = E\frac{zf}{s^2CF} + \frac{c_2B(z+q)}{bwq^2} - \frac{1}{\cos wjk^2}(k-je) \tag{3.10}$$

The complex interdependencies between $\delta.\Delta \in b$ the many design and performance factors $Ezf/(s^2CF)$ are shown by equation 3.10. To handle these intricate interrelationships $c_2B(z+q)$, the ADIOS framework TRIZ and scalability computingbwq^2, which enables systemic optimization of net catchers for electromagnetic emissions $1/(coswjk^2)$. Improved design efficiency, strong performance $(k - je)$, and compliance with strict regulatory criteria.

$$\frac{d^2r}{drc_2} = \left(ur^2 + \frac{\partial f}{1-\forall(r-\rho\tau)}\right) + \frac{(q-p)+py(r-s)}{\tan(q-p)+1} - \frac{1}{\tan(\delta-\beta)} \tag{3.11}$$

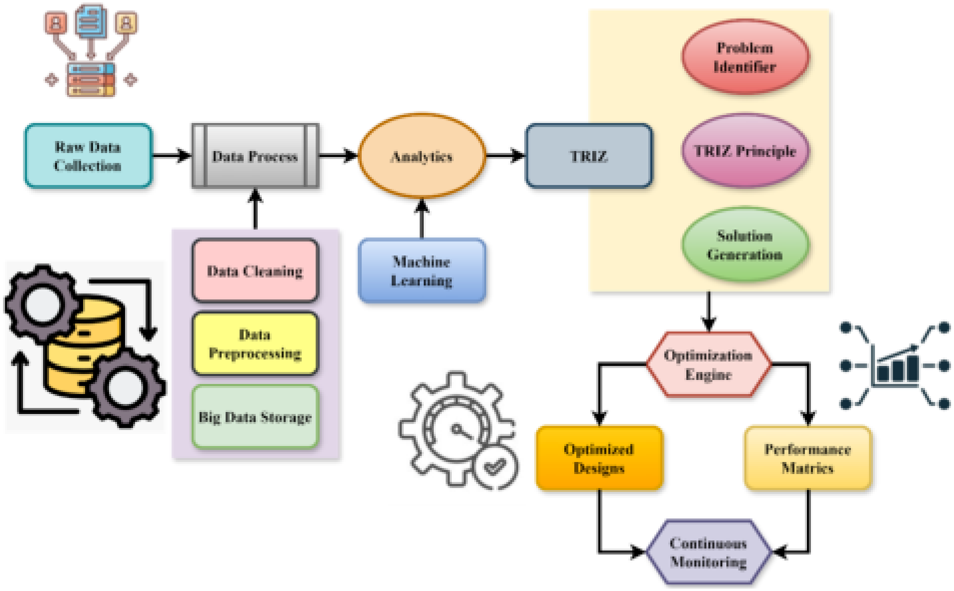


Fig. 3.4: TRIZ theory in optimization systems

In the ADIOS framework $\frac{d^2r}{drc_2}$, the optimization process is characterized by complex linkages $(q-p)+py(r-s)$ and dependencies $\tan(q-p)+1$, as shown in equation (11). The ADIOS system optimizes the design of electromagnetic emission net catchers $\frac{\partial f}{1-\forall(r-\rho\tau)}$ by efficiently managing these complicated interactions $\tan(\delta-\beta)$ techniques and scalable computing.

$$\frac{e^2z}{rs^2q} = \left(t^2 + \frac{f\alpha}{1-(w+sp)}\right) + \frac{(1-ph)}{1+\tan(q-r)} - \frac{1}{m-tp} \tag{3.12}$$

When optimizing microwave emission net catchers, the many interdependencies $\frac{e^2z}{rs^2q}$ and interactions $\frac{f\alpha}{1-(w+sp)}$ are captured by equation 3.12. The ADIOS framework optimizes $(1-ph)$ and analyzes these complex interactions $(1-ph)$ using scalable computation and TRIZ approaches. The ADIOS system increases performance and allows for more precise design alterations by controlling variables including electromagnetic characteristics $\frac{1}{m-tp}$.

Using analytics, raw data collecting, and the TRIZ approach, the figure3.4 shows a complete framework for improving design processes is shown in figure 3.4. Data cleansing, preparation, and storage in a big data system are the first steps in processing and processing raw data. To solve design difficulties in a methodical way, several stages are followed: identify problems, use TRIZ principles to generate inventive ideas, and finally, generate successful solutions. An optimization engine is used to verify and further enhance the design based on the improved solutions. Using performance measures, the framework guarantees constant monitoring and assessment, enabling continual improvement and adaptability. Industries such as manufacturing, engineering, and product development may greatly benefit from this integrated approach, which combines data-driven insights with TRIZ principles. It boosts the efficiency and efficacy of design processes. The product is a set of optimized designs that are both reliable and perform to expectations.

$$bws \mp ewq = 8fip + \frac{1}{2}([su-hp]) - \sin + \frac{1}{2}(u+\infty w) \tag{3.13}$$

When it comes to assessing bws and design parameters ewq, ADIOS effectively manages the computational complexity$8fjp$ by using scalable computing $[su-hp]$. With this method, electromagnetic emission net catchers

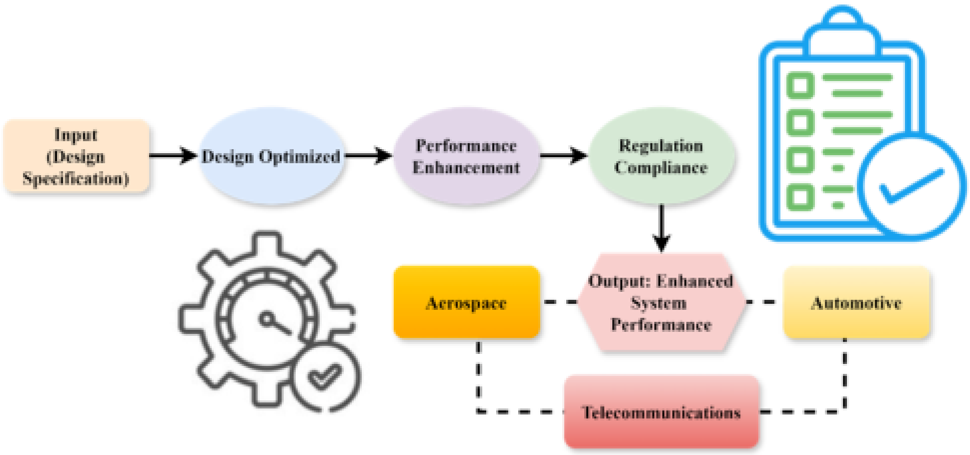


Fig. 3.5: Design specification phase in critical industries

are fine-tuned for dependability sin, performance $(u + \infty w)$, and regulatory compliance by detailed modeling and incremental refinement.

$$\int_{g}^{(h-w)} (1 + \forall w) + jY(p-q) = \int_{Fs}^{2} wp - th(p-1) \tag{3.14}$$

The complex interplay of integrals including factors that impact the optimization $1 + \forall w$ and construction of electromagnetic emission net catchers $jY(p-q)$ is shown in Equation 14. Such equations are methodically $(h-w)$examined to maximize performance p-1 inside the ADIOS framework, which combines TRIZ approaches with scalable computing $wp-th$. Finding and fixing the capsule heater's design flaws is doing a needs assessment based on the ADIOS principle. consider user needs and satisfaction while doing the product designs and features how it is used, the processes of human-machine interaction, and the environment in which the product is used is explained.

Objective 3: Application and Validation in Critical Industries. A well-defined set of steps, the ADIOS framework optimizes electromagnetic emission net catchers more efficiently and effectively. The process starts with entering design requirements, which serve as the basis for all that follows. At the very beginning of the ADIOS architecture, at the Design Optimization phase, it use sophisticated algorithms and TRIZ concepts to hone the preliminary designs. The framework of the next step involves simulation and analysis. In this, optimized designs are extensively tested and analyzed through multimedia to identify areas for improvement and guarantee dependable performance under different conditions. This is followed by Performance Enhancement which further polishes the designs based on the outcomes of simulations for maximum efficiency as well as usefulness. A critical part of enhancing system dependability is EMI Management, which deals with and reduces problems related to electromagnetic interference. The next step is to verify the Regulatory Compliance of these designs to ensure that they adhere to all the rules and regulations. Enhanced System Performance results in highly efficient systems that are also very reliable as per strict regulatory requirements in key industries such as aerospace, telecommunications, automotive among others.

$$f(d-p) = \frac{1}{mje} - \sum_{p=1}^{f} \frac{h(p-w)}{q+pj} - (p-1) + Hp(y-fg) \tag{3.15}$$

Equation 3.15 summarizes this complicated connection $f(d-p)$. The computational complexity of assessing $1/mje$ and changing these equations $h(p-w)/(q+pj)$ is managed by ADIOS using scalable computing, which guarantees accurate modeling of design parameters $(p-1)$. Improved design efficiency $Hp(y-fg)$ and performance.

$$f^{h+1} = R_s Q \log\left(1 - \frac{[Q_{r-1}] - A_{m,Q} + R}{P_1 - E(\partial_2 Q)}\right) - S_w(n-1) \tag{3.16}$$

In the Analysis of electronic system performance f^{h+1} the connection represented by equation 3.16 is critical for assessing R_sQlog the performance of electronic systems. The evaluating exponential $[Q_{r-1}] - A_{m.Q} + R$ and differential calculations $P_1 - E(\partial_2 Q)$maybe handled by ADIOS, which ensures the exact modeling of electrical system characteristics $S_w(n-1)$.

$$s_{p+1} = Fd_{r-1} + Q_{wyp} - \left(1 + \frac{[Y] + F_{p-1}}{\sqrt{g+pj}}\right) - \sum_{s=1}^{k}(q-p) \tag{3.17}$$

When improving $s_(p+1)$ electromagnetic emission net catchers, equation 3.17 Analysis of electromagnetic Fd_{r-} interference management for managing electromagnetic interference Q_wyp. Innovative ways to eliminate interference are made possible by ADIOS via the use of TRIZ principles $\left(1 + \frac{[Y]+F_{p-1}}{\sqrt{g+pj}}\right)$, which enhances compatibility with electromagnetic waves $(q-p)$ and system effectiveness.

$$b^f(q-1) = c\sum_{h=1}^{f} \frac{Q_{s+1}}{\alpha^2} - \frac{(m-1) - Q^{k-1}}{w+qp} - h_{jp}(q-1) \tag{3.18}$$

Analysis of efficiency relies heavily on equation 18, which is of particular importance for improving net catchers that capture electromagnetic emissions $b^f(q-1)$. To guarantee accurate modeling of efficiency factors and their interactions Q_{s+1}/α^2 , ADIOS can manage the complicated calculations required to evaluate sums $(m-1)-Q^{k-1}$ and fractions by using scalable computing $w+qp$. Optimizing design factors including material utilization, electromagnetic shielding effectiveness $h_{jp}(q-1)$.

$$c_q^{r-1} = \frac{E_r}{Q_p - 1} + (1 - \alpha w) - b_w + M - \frac{M}{nW}(p+1) \tag{3.19}$$

When optimizing equipment like electromagnetic emission net catchers c_q^{r-1}, Analysis of the scalability of Scalable Computing$E_r/(Q_p-1)$ may be evaluated with the use of equation 3.19, which is crucial in this context $(1-\alpha w)$. For precise modeling of scalability variables $b_w + M$, ADIOS's scalable computing capabilities are essential for effectively handling the computational needs of assessing and changing$M/nW(p+1)$.

$$D_q = \frac{h}{n-p}(a-q) + c^{wr} - \frac{[1+p] - (y_{jkp} - nm^t)}{pm^k} \tag{3.20}$$

Particularly in the optimization of electromagnetic emission net catchers, equation 3.20 is used for the analysis of design processes D_q. Analysis of design processes is determined with this $h/(n-p)(a-q)$. To improve design processes c^wr, ADIOS applies TRIZ concepts to help find new solutions that balance issues including electromagnetic compatibility $[1+p] - (y_{jkp} - nm^t)$, structural configuration $p + m^k$, and material selection.

Through iterative enhancement of initial design parameters running comprehensive simulations fixing EMI issues, and checking conformity with regulations ADIOS enhances system performance. It can improve designs holistically within industries like aerospace, telecommunications and automotive leading to efficient and reliable systems.

4. Result and discussion. Thus it is necessary to have a comprehensive assessment of electronic system operation considering EMI control for effectiveness improvement or in terms of regulatory compliance. Automated Decision Inspection Optimization System (ADIOS) presents an effective way out for complex design problems by combining TRIZ theory together with scalable computing power. ADIOS leverages big data analytics and machine learning algorithms that enable quick processing and evaluation of huge amounts of information for purposes finding faults in electromagnetic emission net catchers. This method's increased system performance less EMI better reliability has various application sectors include but not limited from Aerospace, Telecommunications and Automotive.

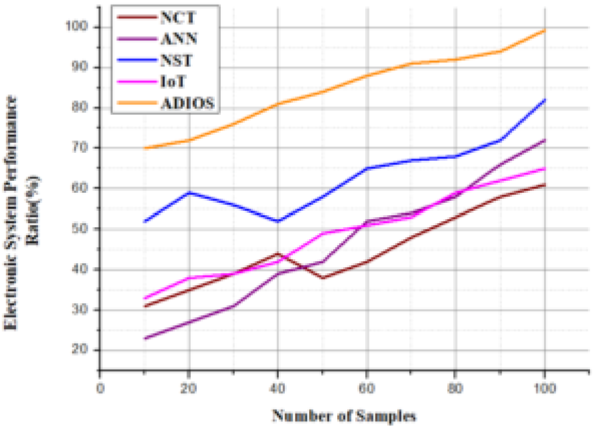


Fig. 4.1: Graph of electronic system performance

4.1. Dataset Description. When trying to maximize or minimize a linear function with several choice variables and constraints, optimization is a common tool in operations research. This makes it useful for issues like scheduling, transportation network design, warehouse site allocation, and production planning. In this case, drew on a consulting engagement in which it advised one of our portfolio firms on choosing a cellular provider that would cost them the least amount of money while still meeting all their needs (total number of lines and pooled data quantity) [26].

The system's accuracy and flexibility are confirmed by controlled laboratory experiments that evaluate its performance under particular, quantifiable situations. To measure how well ADIOS performs, it is also compared to other approaches that have already been developed and implemented, such as the Internet of Things (IoT),Artificial Neural Networks (ANN), Numerical Computational Techniques (NCT), and Nano Sensor Technology (NST).

4.2. Analysis of Electronic System Performance. Improving efficiency and staying in line with regulations both depend on conducting thorough analyses of electronic system performance within the framework of EMI control. The ADIOS offers a powerful method for handling difficult design problems by combining TRIZ theory with scalable computers is explained in equation 16. Issues with electromagnetic emission net catchers may be found and fixed with the help of ADIOS's fast processing and analysis of enormous datasets made possible by machine learning, analytics, and big data. Within the ADIOS framework's simulation tests reveal that it can enhance electronic systems' performance by reducing electromagnetic interference thereby resulting into more efficient and reliable systems. For example in industries like aircrafts, telecommunication sector or car industry; this optimization approach ensures high quality performances while still meeting stringent regulatory requirements. From these results it appears that ADIOS could be used as a tool for improving the EMI management capabilities of electronic systems through new features and modifications to existing design parameters. A 99.25% improvement of the proposed method in electronic system performance ADIOS is shown in figure 4.1.

4.3. Analysis of Electromagnetic Interference Management. To determine that these electronic systems are secure and reliable it is important to analyze EMI control. Effective EMI management aims at identifying undesired electromagnetic emissions that affect system performance. One of the ways to approach EMI problems is through use of modern technologies such as ADIOS which merges TRIZ theory with scalable computers. For instance, by allowing detailed simulation and study on electromagnetic interactions, ADIOS helps identify interference sources and optimal design parameters as done in equation 17. Using distributed computing, machine learning and big data, ADIOS can manage massive computational workloads effectively while providing for effective solutions in terms of EMI management. Major beneficiaries from this system include aerospace, telecommunications, automotive among others which have high stakes and stringent regulatory

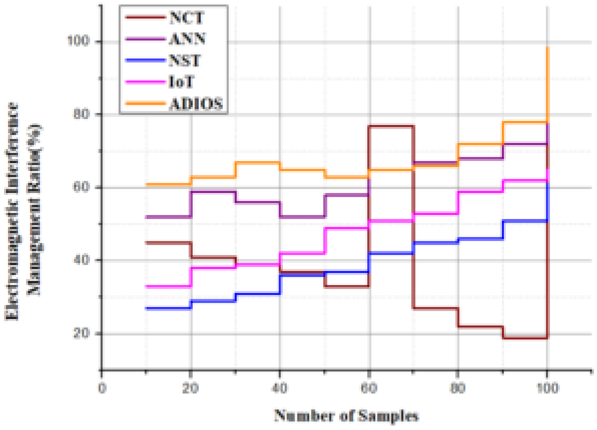


Fig. 4.2: Graph representation of electromagnetic interference management

expectations. Therefore, through accurate analysis and inventive optimization ADIOS improves EMI control enabling electronic systems to operate without interferences and optimum efficiency across various fields..In the proposed method of ADIOS the electromagnetic interference management ratio is improved by 98.41% is shown in figure 4.2.

4.4. Analysis of Efficiency. To maximize the effectiveness and use of resources, efficiency analysis is crucial. Efficiency analysis finds ways to increase production and decrease waste by looking at different parts of the business. Assessing the design, material integration, and electromagnetic interactions of electromagnetic emission net catchers is an important part of efficiency analysis is derived in equation 18. The goal is to maximize performance while minimizing energy loss. An effective method for analysing efficiency is the ADIOS, which combines TRIZ theory with scalable computing. By combining big data, machine learning, and distributed computing, ADIOS can handle massive datasets, model complicated situations, and pinpoint inefficiencies with pinpoint accuracy. The total performance of the system is improved by this all-encompassing method, which enables focused optimization. These findings are useful for sectors like aviation, telecoms, and automobiles since more efficiency means lower operating costs, more regulatory compliance, and more efficient and reliable systems. In figure 4.3, the efficiency ratio is gradually increased by 98.21% in the existing method.

4.5. Analysis of Scalability of Scalable Computing. Determining a system's capacity to manage growing workloads or increase its resources without sacrificing performance is the primary goal of scalability analysis in scalable computing is shown in figure 4.4. Scalable computing comes in to process large data sets and complex simulations by utilizing distributed computing infrastructures. Scalability is therefore vital for optimization of electromagnetic emission net catchers plus other systems that have escalating computational requirements. The equation 19 describes one such system, the ADIOS. Scalability analysis involves monitoring how well the system performs with additional computing resources. This examination also includes some considerations like load balancing and parallel process efficiency as well as resource allocation. High performance, reliability and efficiency as the system scales up calls for a comprehensive scalability study. Therefore, this is highly valued in industries with heavy computational workload like aerospace, telecoms and automobile industry among others. For instance; scalable computing technologies such as ADIOS enhance system performance and capabilities through a strong scalability that enables them to deliver effective solutions regardless of demand size since they are capable of scaling up or down when required. In the proposed method the previous one had a ratio of scalable computing which was increased by 96.31%.

4.6. Analysis of Design Processes. To improve and optimize the development of complicated systems, it is vital to analyse the design processes. This requires looking for inefficiencies and ways to improve at every step of the design process, from coming up with the idea to putting it into action. Complete consideration

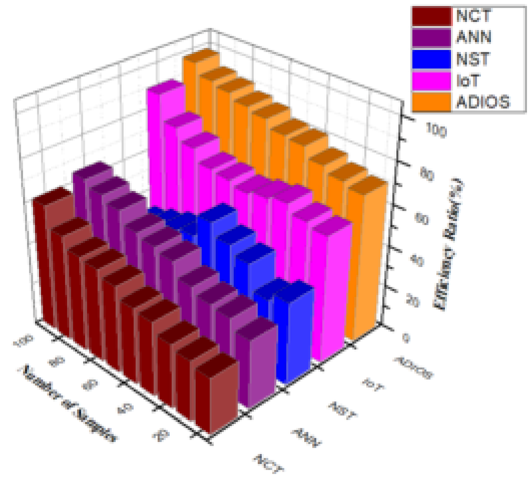


Fig. 4.3: Graph representation of efficiency

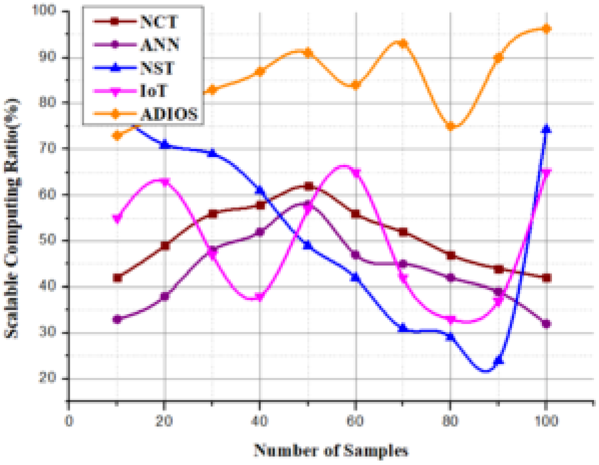


Fig. 4.4: Graphical representation of scalable computing

of all criteria, such as electromagnetic interactions, material selection, and structural integrity, is achieved by comprehensive design process analysis in the context of electromagnetic emission net catchers is explained in equation 20. The ADIOS is one example of a cutting-edge technology that combines TRIZ theory with scalable computers to provide a systematic approach to improving design processes. To maximize performance and conform to regulatory requirements, ADIOS analyses and improves design aspects using big data, machine learning, and comprehensive simulations. Improved system efficiency, reliability, and innovation may be achieved with the help of ADIOS by methodically tackling design difficulties and repeatedly testing solutions. Industries like aerospace, telecommunications, and automotive greatly benefit from an all-encompassing approach because of the clear correlation between the effectiveness of a system and its design procedures. The design processes ratio is improved by 96.24% in the proposed method of ADIOS is shown in figure 4.5.

Simulation studies reveal substantial gains in electronic system performance due to reductions in electromagnetic interference, proving the efficacy and scalability of the ADIOS architecture. ADIOS is to optimize design parameters and implement creative solutions so that electronic systems may run efficiently and with little disturbance, following all regulatory criteria. The framework has great promise as an effective instrument

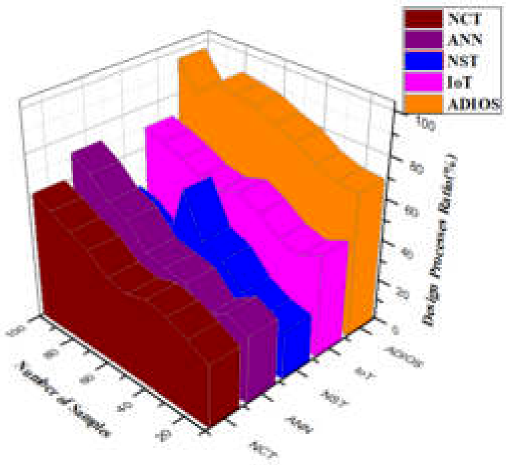


Fig. 4.5: Graph of design processes

for EMI control due to its capacity to handle massive computational workloads while producing accurate results. ADIOS hasincreased efficiency, scalability, and regulatory compliance in high-stakes sectors while providing a complete electronic system design and performance solution.

The suggested Automated Decision Inspection Optimization System (ADIOS) considers environmental factors for improved efficiency and resilience in optimizing electromagnetic emission net catchers. For ADIOS to tailor its decision-making to its operational environment, it incorporates real-time data on humidity, temperature, electromagnetic interference, and topography. The system can adapt its optimization tactics on the fly, guaranteeing stable performance regardless of obstacles or environmental changes. Suppose there's a rise in humidity, for instance. In that case, ADIOS may change its settings to ensure that electromagnetic signals travel through the air more efficiently, or it can compensate for electromagnetic interference caused by neighbouring electronics. In real-world applications, where situations are often varied and unexpected, ADIOS is more suited since it considers these environmental elements, improving its dependability and efficiency.

5. Conclusion. The main emphasis of this paper is on the current modular design method's interface structure design difficulty. Incorporating the TRIZ and EME integration models, it enhances the current modular design approach. After doing an EMI analysis of the interface coupling to the functional requirements, this technique decouples the two and suggests a good connection topology for the modular product using the TRIZ conflict solution tool. By summarizing the basic engineering parameters often used in modular design and analysing the needs and conflicts in the module structure design, this article makes it easy to locate appropriate parameters. When it comes time to divide modules according to TRIZ theory, the ADIOS parameters might be useful supplementary tools for identifying the principle (technical)correlation between sections. The high configuration design, which relies on TRIZ and EMI, proved the practicability of the modular design approach. Innovative methods to optimizing the design of electromagnetic devices, particularly in relation to the use of ADIOS.

The rapid advancement of sophisticated TRIZ algorithms and intelligent manufacturing technologies presents both possibilities and problems for the optimization of electromagnetic device designs, as discussed at this meeting. Optimal design of electromagnetic devices presents several difficult challenges, not the least of which is meeting performance requirements while simultaneously achieving high dependability, robustness, manufacturing quality, and adaptability during the device's lifespan. It remains challenging to discover the link between possible functional needs and design parameters due to the absence of methodological direction and some subjectivity throughout the process, beginning with the module division and continuing through the design matrix for unit modules. Consequently, more research into methods for efficiently identifying issues via the establishment of module interactions and the facilitation of the design matrix is required to enhance product design efficiency. The proposed method increases the Electronic System Performance ratio of 99.25%, Electromagnetic Interference Management ratio of 98.41%, Efficiency ratio of 98.21%, Scalable Computing ratio of 96.31%, and Design Processes ratio of 96.24% compared to existing methods.

Acknowledgement. Wuhan Social Science Federation Project – Research on the Integration and Development of Culture and Technology.

REFERENCES

[1] Perricone, V., Santulli, C., Rendina, F., & Langella, C. (2021). Organismal design and biomimetics: a problem of scale. Biomimetics, 6(4), 56.

[2] Wang, R., Milisavljevic-Syed, J., Guo, L., Huang, Y., & Wang, G. (2021). Knowledge-based design guidance system for cloud-based decision support in the design of complex engineered systems. Journal of Mechanical Design, 143(7), 072001.

[3] Petutschnig, L., Rome, E., Lückerath, D., Milde, K., Gerger Swartling, Å., Aall, C., & Kienberger, S. (2023). Research advancements for impact chain-based climate risk and vulnerability assessments. Frontiers in Climate, 5, 1095631.

[4] Flapper, R. A. C., Huijben, J. C. C. M., Bobelyn, A. S. A., Wiegmann, P. M., & Sickert, L. (2023). Improving sustainability through Product Lifecycle Management in the Packaging Industry; a Tetra Pak case study.

[5] Ngeoywijit, S., Kruasom, T., Ugsornwongand, K., Pitakaso, R., Sirirak, W., Nanthasamroeng, N., ... & Kaewta, C. (2022). Open innovations for tourism logistics design: a case study of a smart bus route design for the medical tourist in the City of Greater Mekong Subregion. Journal of Open Innovation: Technology, Market, and Complexity, 8(4), 173.

[6] Lenau, T. A., & Lakhtakia, A. (2022). Biologically Inspired Design: A Primer. Springer Nature.

[7] Catalano, G. D., & Catalano, K. C. (2020). Engineering Design: An Organic Approach to Solving Complex Problems in the Modern World. Morgan & Claypool Publishers.

[8] Bianciardi, A., & Cascini, G. (2022). A bio-inspired approach for boosting innovation in the separation technology sector. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 236(9), 4533-4550.

[9] Altalhi, T. (Ed.). (2023). Green Sustainable Process for Chemical and Environmental Engineering and Science: Carbon Dioxide Capture and Utilization. Elsevier.

[10] Gomez, P., & Lambertz, M. (2023). Leading by Weak Signals: Using Small Data to Master Complexity (Vol. 5). Walter de Gruyter GmbH & Co KG.

[11] Kaliteevskii, V., Bryksin, M., & Chechurin, L. (2021). TRIZ application for digital product design and management. In Creative Solutions for a Sustainable Development: 21st International TRIZ Future Conference, TFC 2021, Bolzano, Italy, September 22–24, 2021, Proceedings 21 (pp. 245-255). Springer International Publishing.

[12] Chang, D., Li, F., Xue, J., & Zhang, L. (2023). A TRIZ-inspired knowledge-driven approach for user-centric smart product-service system: A case study on intelligent test tube rack design. Advanced engineering informatics, 56, 101901.

[13] Wang, C. N., Nhieu, N. L., & Viet, T. A. P. (2024). Enhancing efficiency in PCB assembly for the leading global electronics manufacturing services firm: a TRIZ and Ant Colony Optimization approach. The International Journal of Advanced Manufacturing Technology, 1-24.

[14] Kaliteevskii, V., Bryksin, M., & Chechurin, L. (2022, September). Integration of TRIZ methodologies into the digital product development process. In International TRIZ future conference (pp. 273-284). Cham: Springer International Publishing.

[15] Aguilar-Lasserre, A. A., Torres-Sánchez, V. E., Fernández-Lambert, G., Azzaro-Pantel, C., Cortes-Robles, G., & Román-del Valle, M. A. (2020). Functional optimization of a Persian lime packing using TRIZ and multi-objective genetic algorithms. Computers & Industrial Engineering, 139, 105558.

[16] Brad, S. (2021). Domain analysis with TRIZ to define an effective "Design for Excellence" framework. In Creative Solutions for a Sustainable Development: 21st International TRIZ Future Conference, TFC 2021, Bolzano, Italy, September 22–24, 2021, Proceedings 21 (pp. 426-444). Springer International Publishing.

[17] Delgado-Maciel, J., Cortés-Robles, G., Sánchez-Ramírez, C., García-Alcaraz, J., & Méndez-Contreras, J. M. (2020). The evaluation of conceptual design through dynamic simulation: A proposal based on TRIZ and system Dynamics. Computers & Industrial Engineering, 149, 106785.

[18] Basuki, A., Cahyani, A. D., & Umam, F. (2024). Application of the Triz Model for Evaluating the Potential Innovation Value of a Digital Start-Up Company. Management Systems in Production Engineering, 32(2), 202-211.

[19] Al-Betar, M. A., Alomari, O. A., & Abu-Romman, S. M. (2020). A TRIZ-inspired bat algorithm for gene selection in cancer classification. Genomics, 112(1), 114-126.

[20] Ren, S., Gui, F., Zhao, Y., Zhan, M., Wang, W., & Zhou, J. (2021). An extenics-based scheduled configuration methodology for low-carbon product design in consideration of contradictory problem solving. Sustainability, 13(11), 5859.

[21] Gunasekaran, K., Kumar, V. V., Kaladevi, A. C., Mahesh, T. R., Bhat, C. R., & Venkatesan, K. (2023). Smart Decision-Making and Communication Strategy in Industrial Internet of Things. IEEE Access, 11, 28222–28235.

[22] https://doi.org/10.1109/access.2023.3258407.

[23] Kumar, V., Niveditha, V. R., Muthukumaran, V., Kumar, S. S., Kumta, S. D., & R., M. (2021). A Quantum Technology-Based LiFi Security Using Quantum Key Distribution. Advances in Computational Intelligence and Robotics, 104–116. https://doi.org/10.4018/978-1-7998-6870-5.ch007.

[24] Umamaheswaran, S., Lakshmanan, R., Vinothkumar, V., Arvind, K. S., & Nagarajan, S. (2019). New and robust composite micro structure descriptor (CMSD) for CBIR. International Journal of Speech Technology, 23(2), 243–249.

[25] https://doi.org/10.1007/s10772-019-09663-0.

[26] Kumar, Dr. V. V., Arvind, Dr. K. S., Umamaheswaran, Dr. S., & Suganya, K. S. (2019). Hierarchal Trust Certificate Distribution using Distributed CA in MANET. International Journal of Innovative Technology and Exploring Engineering, 8(10), 2521–2524.

[27] Kumar V., V., Ramamoorthy S., Kumar V., D., Prabu M., & Balajee J. M. (2021). Design and Evaluation of Wi-Fi Offloading Mechanism in Heterogeneous Networks. International Journal of E-Collaboration, 17(1), 60–70. https://doi.org/10.4018/ijec.2021010104.

[28] Kaggle, Code, https://www.kaggle.com/code/dhitology/optimization

Edited by: Dhilip Kumar V
Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing
Received: Jul 25, 2024
Accepted: Oct 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1212–1221, DOI 10.12694/scpe.v26i3.4080

CONSTRUCTION OF A POWER MARKET TRADING PLATFORM BASED ON REGIONAL BLOCKCHAIN TECHNOLOGY

HONGXI WANG[*], XUDONG ZHANG[†], FEI LI [‡], LUN SHI [§], YIDI WU [¶], AND CHUNHAI LI[‖]

Abstract. In order to solve the problems of traditional centralized trading platforms being difficult to handle rapidly increasing transaction data, achieving cross regional information sharing and resource unified optimization configuration, the author proposes the construction of a power market trading platform based on regional blockchain technology. This technology is based on the regional power energy trading architecture, designs a decentralized cloud energy storage blockchain trading model, and further refines the three-layer technical architecture of the business layer, middleware, and open license chain, thereby ensuring effective collaboration between the client and the distributed backend. In order to further improve the computational efficiency of the system, the author proposes a consensus algorithm optimization scheme based on transaction credit evaluation. The evaluation of blockchain nodes is achieved through the joint evaluation of multidimensional indicators, and the credit ranking of nodes is completed based on the defined transaction priority weight. The experimental results indicate that: Compared with the comparison algorithm, the clustering accuracy obtained by the proposed algorithm is higher, with a maximum value of 91.56%, indicating that the proposed algorithm correctly clusters more electricity sales information. The algorithm proposed by the author can reduce the probability of malicious transactions compared to existing algorithms, while improving the processing power and response speed of the trading system.

Key words: Blockchain, Electric energy, Credit evaluation, Consensus algorithm

1. Introduction.

In a narrow sense, the electricity market refers to the mechanism by which electricity producers and users determine prices and quantities through competition. The construction of the electricity market and the establishment of trading platforms are of utmost importance in the reform of the electricity system, and are also key factors in the success or failure of the reform [1]. The electricity market trading platform is a technical support system that serves market operations. The power trading platform is one of the key components of the national power market trading operation system, which contains a wide range of information types and involves a wide range of aspects. Therefore, it has received high attention from the state, regulatory authorities, and various sectors of society. The rapid increase in electricity trading information poses great challenges to the processing and application of electricity trading information [2]. As a distributed shared database technology, blockchain technology has the characteristics of decentralization, transparency, fairness and other characteristics consistent with the concept of the energy Internet. It can be a pattern and mode under the Internet, operate efficiently under the premise of ensuring trust, promoting transactions, achieving authentication and other advantages, and make up for the shortcomings of high transaction costs, asymmetric transaction information, low transaction data efficiency and data security in the traditional power trading mechanism, laying the foundation for building a multi-agent form of energy interconnection, information and physical integration of power trading platform [3].

With the continuous deepening of electricity reform, the market operation will become more complex and variable. In order to better avoid risks, it is important to comprehensively grasp the information trading business of the electricity market, analyze the operation of the electricity market, and timely predict market trends in future electricity trading work. The strengthening of electricity trading market analysis business is conducive to improving service quality and meeting the needs of the company's business development [4]. By collecting

[*] State Grid Hebei Markting Service Center, Hebei Shijiazhuang, 050021, China. (HongxiWang9@126.com)
[†] State Grid Hebei Electric Power Co., Ltd, Hebei Shijiazhuang, 050021, China. (XudongZhang36@163.com)
[‡] State Grid Hebei Markting Service Center, Hebei Shijiazhuang, 050021, China. (FeiLi825@126.com)
[§] State Grid Hebei Markting Service Center, Hebei Shijiazhuang, 050021, China. (LunShi8971@163.com)
[¶] State Grid Hebei Electric Power Co., Ltd, Hebei Shijiazhuang, 050021, China. (YidiWu38@126.com)
[‖] Shijiazhuang Kelin Electric Co.,Ltd, Hebei Shijiazhuang, 050000, China. (Corresponding author, ChunhaiLi8@163.com)

real-time data information, analyzing big data, and monitoring trading business throughout the process, we can grasp market dynamics, conduct early warning analysis, and ensure the healthy and orderly operation of the trading market. Therefore, the reconstruction design of the four major applications of commodity trading, market settlement, market services, and market analysis is of great significance, which can effectively avoid affecting the normal operation of other applications when one application malfunctions. In this way, the stability and security of the trading platform system will be greatly improved [5]. In recent years, the energy Internet, which is highly integrated with new energy power generation and information technology, has provided possible solutions for large-scale utilization and flexible access of all kinds of energy [6]. More and more market entities are transforming from traditional single energy consumers or producers to electricity producers and consumers with independent decision-making abilities, participating in electricity market competition in a more flexible and diverse way. In this new situation, how to build a reasonable and effective distribution side electricity market trading platform, fully guarantee the different interests of various market entities and the effective allocation of energy resources, is currently a key issue that urgently needs to be solved.

2. Literature Review. With the development of digitalization, networking, and information technology, as well as the expansion of the number of users on electricity market trading platforms, there are also higher requirements in terms of computing [7]. The power trading platform, based on the ”cloud” technology architecture and utilizing big data to analyze basic business and expand service channels through mobile applications, has emerged in response to the trend [8]. Cloud services have high security, and transaction rules in different regions can be subdivided into various ”transaction microservices” according to their respective situations. When rules change, only small-scale upgrades and adjustments to microservices can quickly adapt to the rapid development needs of the market. In the future, network technology will continue to mature, and cloud based trading platforms can be continuously upgraded to quickly expand the demand for software and hardware resources, improve the stability, reliability, and efficiency of trading platforms, and ensure that all types of market members can participate in market transactions fairly [9].

Yang et al. introduced an assessment framework for gauging the maturity of blockchain technology applications. This system encompasses five primary indicators: critical application prerequisites, data safeguarding measures, process intricacies, ecosystem coherence, and technical performance benchmarks, each accompanied by relevant secondary indicators [10]. Afzal, M. et al. categorized blockchain technology, consensus algorithms, and smart contract varieties within the context of peer-to-peer energy markets. Their work offers insights into how blockchain can revolutionize conventional markets into advanced iterations. Furthermore, they compiled a range of objective functions and strategies utilized to attain optimal objectives in electricity trading markets, serving as valuable references for researchers delving into blockchain applications in energy market trading [11]. Chen et al. argue that power energy systems are evolving towards greater decentralization, posing challenges to centralized management due to the potential absence or lack of trust in central institutions. They contend that blockchain technology presents a viable solution to this issue by facilitating trusted collaboration among stakeholders without relying on a central authority [12]. Hasan, M. et al. outlined the primary security concerns addressable by big data and blockchain technologies within smart grid contexts. Subsequently, they conducted a comprehensive review of recent blockchain-focused research across diverse literature sources, analyzing their implications for enhancing the security of smart grid systems [13].

The author combines the blockchain distributed accounting framework with the smart contract model for electricity trading to establish an electricity trading platform in a distribution side trading framework with multiple market entities. On the basis of meeting the interests of various trading entities, build a peer-to-peer distributed sharing network architecture to achieve trust and security in transactions. The effectiveness of the proposed power trading platform has been verified through examples, which can provide reference for the application of blockchain technology in the construction of a distribution side power market trading platform.

3. Method.

3.1. System Model. The regional coordination of power supply and demand is constrained by various supply and demand relationships, including policy guidance, institutional construction, platform support, and safety assurance [14]. The architecture of the regional power energy cloud storage trading system is shown in Figure 3.1.

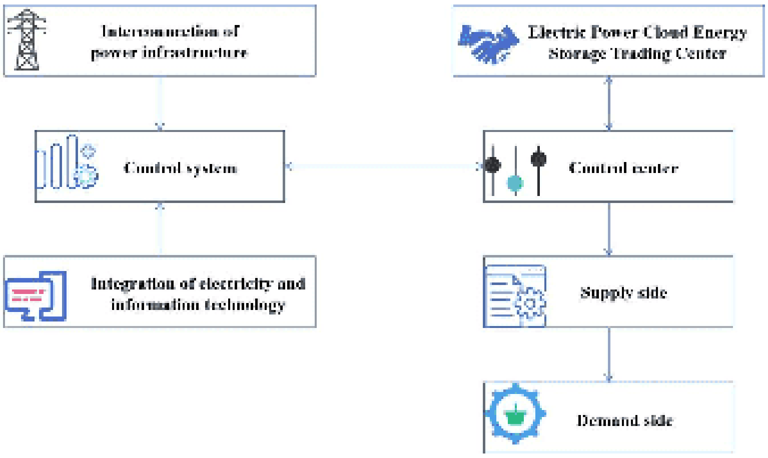

Interconnection of power infrastructure
Electric Power Cloud Energy Storage Trading Center
Control system
Control center
Integration of electricity and information technology
Supply side
Demand side

Fig. 3.1: Architecture of Regional Power Energy Cloud Storage Trading System

The power control center is the dispatch center of the entire system, which interacts with the power control system and the power cloud energy storage trading center. The control system is mainly responsible for receiving power infrastructure interconnection information, achieving the integration of power and information technology, and completing the construction of the power dispatch basic platform. The power cloud energy storage trading center is responsible for completing the transaction scheduling of online power storage units, and sending the transaction information and scheduling data to the control center, which then centrally schedules resources, and then transmits them to the power demand side through the power resource supply side.

3.2. Cloud energy storage transaction blockchain technology architecture. The transaction of the power cloud energy storage trading center needs to ensure the security of the power system, consider the convenience of transactions, and also consider the credibility of the platform. Based on this, the author proposes a blockchain based regional power energy cloud storage trading technology. Blockchain is a decentralized distributed ledger technology that has the characteristics of decentralization and immutability [15]. The overall architecture of power energy trading based on blockchain technology is shown in Figure 3.2.

The blockchain architecture adopted by the author is shown in Figure 3.3, which is mainly divided into three modules: business layer, middleware, and open license chain [16]. Among them, the business layer includes business systems, HSM services, and browsers; Middleware is mainly divided into Application Programming Interface (API), Message Queuing, and Data Processing modules; The open license chain includes communication modules, consensus modules, encryption and signature verification modules, smart contracts, and blockchain data ledgers.

The block structure adopted by the author is shown in Figure 3.3, which mainly includes two parts: Block head and block body. Different blocks can be connected together to form a blockchain.

The block header is mainly responsible for storing connection information between blocks, that is, storing the block number, the parent hash hash value of the block, the hash hash value, timestamp, difficulty target, and random number of the block. The information stored in the block header is used to ensure that the blocks can be orderly connected to the blockchain. The block body stores a large amount of transaction information, and each block is independent, and due to the association of storage parameters with the preceding and following blocks. Therefore, traceability can be carried out and the immutability of block information can be ensured [17].

3.3. Blockchain Cloud Energy Storage Transaction Consensus Technology. The use of blockchain architecture in power energy trading requires ensuring the accuracy and credibility of the entire transaction data. Therefore, it is necessary to use reasonable and efficient consensus algorithms to ensure the ecological balance of system consistency, availability, and partition tolerance. On the basis of practical Byzantine algorithms, the

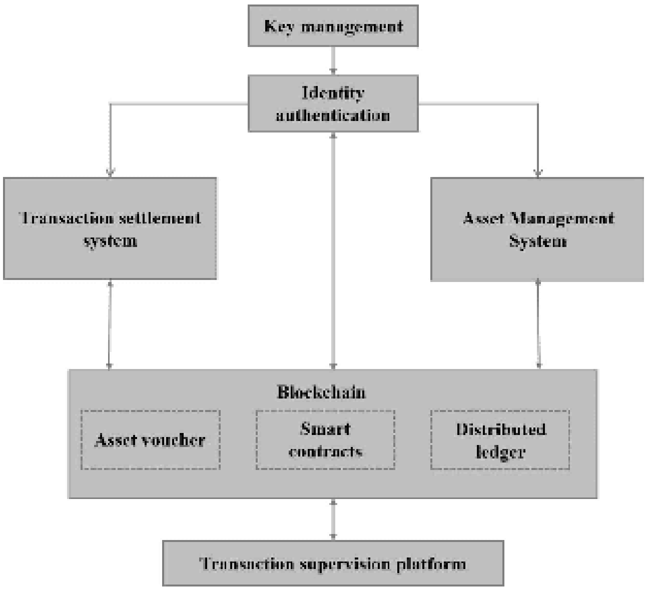


Fig. 3.2: Blockchain Power Cloud Energy Storage Trading Model

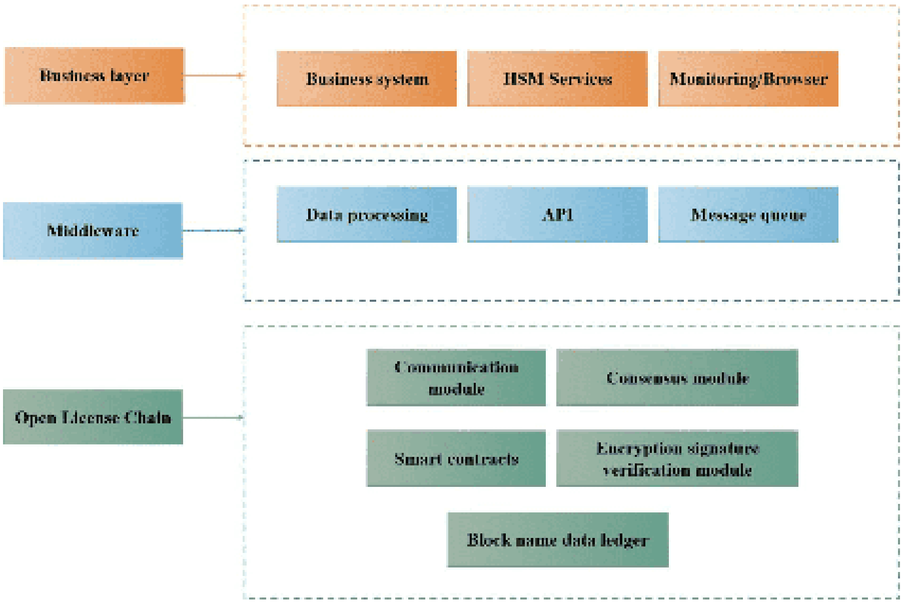


Fig. 3.3: Blockchain Power Cloud Energy Storage Trading Technology Architecture

author evaluates transaction behavior to further suppress malicious nodes and reduce computational power loss, enabling the system to quickly complete information synchronization and reach consensus. The consensus optimization algorithm process based on transaction credit evaluation is shown in Figure 3.5.

The smart contract model for electricity energy trading defined in the article is:

$$B_E = (S, B, C_E, SC, T, \mu) \tag{3.1}$$

In the formula, S represents the power supply side set, B represents the demand side set, C_E represents

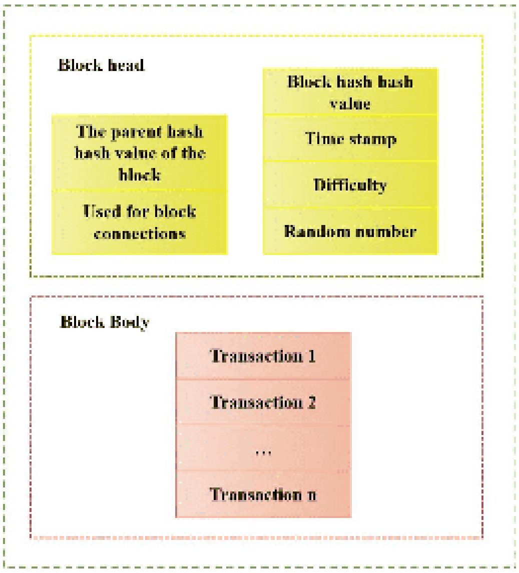


Fig. 3.4: Blockchain Block Structure

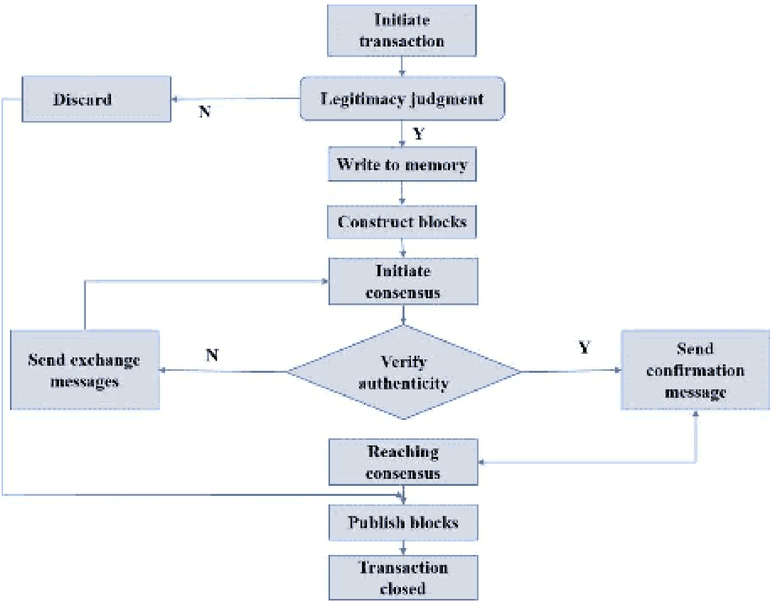


Fig. 3.5: Consensus Optimization Algorithm Process

the transaction blockchain, and SC represents the smart contract, μ represents the mapping between T and C_E. $T = \{t_k | k \in S \times B\}$ represents the set of transactions, and S×B represents the Cartesian product.

The definition of power energy trading blockchain is:

$$C_E = (C_0, C_{0a}) \tag{3.2}$$

Among them, C_0 represents the initial blockchain, and C_{0a} represents the consensus optimization algorithm. The consensus optimization algorithm based on credit evaluation proposed by the author evaluates the accuracy of electricity energy transactions, and its indicators mainly include:

1. C_1: Integrity of power energy dispatch information, including information on the demand side and supply side of power energy, electricity consumption and unit price, etc;
2. C_2: Verify the authenticity of information on the power energy supply side and verify whether the supply side truly exists;
3. C_3: Reasonability of power energy trading time, verified based on timestamp information;
4. C_4: The authenticity of information on the demand side of electricity energy;
5. C_5: Adequacy of funds on the demand side of electricity energy, which can meet the electricity resources required for payment;
6. C_6: Random number matching of transaction information.

In summary, the accuracy evaluation results of electricity energy trading can be expressed as:

$$T_K^n = \prod_{i=1}^{6} C_i \tag{3.3}$$

In the formula, T_K^n represents the accuracy of the transaction, K represents the blockchain number, n represents the transaction number, C_i represents the evaluation indicator, and its quantification is defined as:

$$\begin{cases} C = \{C_i | i = 1, 2, 3, 4, 5, 6\} \\ C_i = 0, false \\ C_i = 1, true \end{cases} \tag{3.4}$$

There are:

$$\begin{cases} T_K^n = 0, invalid \\ T_K^n = 1, effective \end{cases} \tag{3.5}$$

If the transaction is invalid, it indicates that there is an error in the transaction information record. This transaction is illegal and there may be a risk of malicious tampering with transaction information, so blockchain information synchronization should not be carried out.

In order to evaluate the accuracy of node transaction information records, credit value AA is defined as a representation:

$$e_N^l = \frac{\sum_{n=1}^{N_K} T_K^n}{N_K} \tag{3.6}$$

Among them, N_K represents the number of transactions in block K.

In the power trading blockchain, nodes collect transaction information and calculate random number solutions that comply with the blockchain hash function. The solution of this function can be expressed as:

$$H(M, S) \leqslant T_a \times e_N^l = 2^{n-t} e_N^l \tag{3.7}$$

Among them, M represents the Merkle root of the node, S represents the random value of the node's forward solution, and T_a is the system difficulty coefficient.

After hashing the forward solved random numbers and mapping them with a mapping interval of $0 \sim 2^{n-1}$, the probability of obtaining a random number that meets the requirements in a single solution is:

$$P_i^l = \frac{T_a \times e_N^l}{2^n} = \frac{2^{n-t} e_N^l}{2^n} = \lambda e_N^l \tag{3.8}$$

Among them, n represents the mapping space range, t represents the difficulty coefficient, and $\lambda = 2^{-t}$ represents the probability difficulty constant.

Assuming that the computing power of nodes in the blockchain is the same, the transaction credit values of different nodes are different. The probability of node competition winning transaction accounting rights in blockchain can be expressed as:

$$P = \frac{P_i^l}{\sum_{j=1}^{L} P_i^l} = \frac{e_N^l}{\sum_{j=1}^{L} e_N^l} \tag{3.9}$$

In the formula, L represents the number of nodes.

Define the transaction priority value as:

$$V = \frac{c}{e_N^l} \tag{3.10}$$

Among them, c represents the cost of power supply. The higher the credit value, the higher the transaction priority, and the higher the probability of successful node competition. Correspondingly, malicious nodes with low credit values have a lower probability of successful competition.

3.4. Blockchain distributed ledger design. The author designs a distributed ledger for electricity trading based on blockchain technology, which is a database that is shared, replicated, and synchronized among market entities under the blockchain. The main ledger structure includes an account information layer, a security encryption layer, a consensus mechanism layer, and a transaction incentive layer. Based on real account and address information, each market entity node implements legal accounting through encryption technology and consensus mechanisms, and rewards trading participants, so that each node obtains a unique and authentic copy of the ledger, which is tamper proof and transparent and traceable.

3.4.1. Account Information Layer. Each market entity in the blockchain network has a dedicated public key and private key (hexadecimal string). The public key is bound to the true identity and address of the trader and can be publicly released to users across the network, while the private key is only held by the user themselves. The design of blockchain private keys ensures unique ownership under the corresponding account address, where each account address has and only corresponds to one private key. The use of private keys adopts the Hash algorithm, which maps a string to another fixed length string, and the two are not independent. That is, after a series of encryption operations, the private key can obtain the address, but cannot be inferred from the address.

3.4.2. Security encryption layer. Not only do the public and private keys of the account have an encryption relationship, but the transaction bills under each node are also encrypted, reflected in the digital digest and signature of the trader. Digital summarization is a hash operation performed on digital content to obtain a unique string to refer to the original and complete digital content, ensuring that the original content has not been tampered with; Market entities can use private keys to sign summary information. After the electricity selling entity encrypts the transaction bill with a private key, other users verify the authenticity of the data source by decrypting it based on the electricity selling entity's public key, which can be regarded as the inverse operation of the signature process. If the decryption value is consistent with the digital signature in the original transaction bill, it indicates that the transaction is valid and is allowed to be added to the blockchain ledger [18].

3.4.3. Consensus mechanism layer. Blockchain has the characteristics of distribution, autonomy, openness and free access, so there is no central node to ensure the consistency of accounting among each node. The author adopts a proof of work (PoW) mechanism to ensure consensus on transaction bills for blockchain addition. Each market entity continuously adds random numbers (nonce) to new blocks that have not yet joined the chain for hash operations, in order to compete for unique accounting rights in this round by solving cryptographic problems (i.e. proof of workload). Because only one node in each round can successfully account and add the new block information to the blockchain network, other nodes that fail to compete will stop competing for accounting rights, copy the new block information and add it to their own node database, thereby ensuring the uniqueness and authority of the blockchain ledger, enabling all nodes in the network to reach consensus and share data.

Table 4.1: Clustering accuracy data Table

Number of	Accuracy (%)	
experiments/time	Propose an algorithm	Comparison algorithm
1	89.56	65.41
2	78.93	58.51
3	80.34	55.32
4	91.56	59.13
5	85.57	57.32
6	90.26	60.20
7	85.45	72.08

3.4.4. Transaction incentive layer. As the trading volume increases, in order to avoid too many invalid transactions, the trading network continuously increases the difficulty and cost of reaching consensus, that is, the number of digits starting with 0 calculated by Hash continues to increase. Therefore, a reasonable incentive mechanism can be introduced into the consensus mechanism to promote active participation of traders in the development of blockchain, encourage effective accounting behavior, and align the self-interest behavior of consensus nodes to maximize profits with the overall goal of ensuring the security and effectiveness of decentralized systems. In the PoW mechanism, the trading network adds a transfer transaction as a recognition reward to the node that wins the accounting rights.

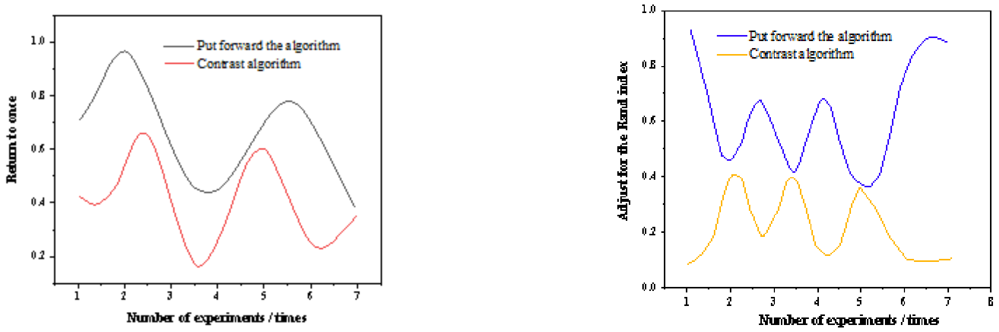
3.5. Experimental Analysis. In order to verify the application performance of the proposed algorithm, a power load data clustering algorithm based on DTW histogram was selected as a comparative algorithm, and a comparative experiment was designed. The specific experimental process is as follows. Selecting the electricity sales information generated by a certain regional electricity trading center as the experimental object, although the selected experimental object is the electricity trading center within a small area, the number of users is still considerable, and the quantity of electricity sales information is still high. For the convenience of conducting experiments, the electricity sales information generated by the electricity trading center in a certain week is used as experimental data to reduce the amount of experimental data and ensure the stability of the experiment [19]. Based on the selected experimental subjects, in order to quantify the clustering effect of the integrated distribution of electricity information, clustering accuracy, normalization degree, and their adjusted Rand index are selected as evaluation indicators. The calculation formula is:

$$\begin{cases} ACC = \frac{N_{cor}}{N} \\ NMI = \frac{2I(X,Y)}{H(X)+H(Y)} \\ ARI = \frac{RI-E[RI]}{MAX[RI]-E[RI]} \end{cases} \tag{3.11}$$

In the formula, ACC, NMI, and ARI represent clustering accuracy, degree of regression, and their adjusted Rand index, respectively; N_{cor} represents the number of correctly clustered electricity sales information; N represents the total number of electricity sales information; I (X, Y) represents the correlation between any two electricity sales information. H(X) and H(Y) represent the maximum entropy of the electricity sales information X and Y, respectively; RI represents clustering coefficient; E [RI] and MAX [RI] represent the clustering coefficient error value and maximum value, respectively.

4. Experimental Results and Discussion. The clustering accuracy data obtained through experiments are shown in Table 4.1. As shown in Table 4.1, compared with the comparative algorithms, the clustering accuracy obtained by the proposed algorithm is higher, with a maximum value of 91.56%, indicating that the proposed algorithm correctly clusters more electricity supply information.

The normalization and adjusted Rand index obtained through experiments are shown in Figure 4.1. As shown in Figure 4.1, compared with the comparison algorithm, the proposed algorithm obtained larger values of normalization and adjusted Rand index, with maximum values reaching 0.98 and 0.92, respectively.



(a) Return to one degree

(b) Adjusting the Rand index

Fig. 4.1: Schematic diagram of normalization and adjustment of Rand index

The clustering accuracy, normalization, and adjusted Rand index obtained by the proposed algorithm are all relatively high, which fully confirms that the integrated clustering effect of the proposed algorithm on electricity sales information is better.

5. Conclusion. The author proposes the construction of a power market trading platform based on regional blockchain technology, conducts in-depth research on blockchain based power energy cloud storage trading technology, designs a blockchain distributed trading architecture, and realizes the module design of four-dimensional integration of trading, scheduling, security, and supervision. In addition, the author further proposed a consensus optimization algorithm based on transaction credit evaluation, which effectively improves system efficiency and fault tolerance. Through multidimensional evaluation and credit ranking, malicious nodes have been effectively suppressed, system processing capabilities have been improved, and response latency has been reduced. Subsequent work can be further optimized to meet the high concurrency requirements of blockchain technology.

REFERENCES

[1] Gao, X., Chan, K. W., Xia, S., Zhang, X., & Wang, G. (2022). Bidding strategy for coordinated operation of wind power plants and ngg-p2g units in electricity market. CSEE Journal of Power and Energy Systems, 8(1), 212-224.

[2] Li, Z., & Ma, T. (2022). Distributed photovoltaics with peer-to-peer electricity trading. Energy and Built Environment, 3(4), 424-432.

[3] Ronaghi, M. H. (2023). A contextualized study of?blockchain technology adoption as a digital currency platform under sanctions. Management Decision, 61(5), 1352-1373.

[4] Santos, D. D., & Gama, P. (2023). Insiders' characteristics and market timing capabilities: buying and selling evidence. Studies in Economics and Finance, 40(2), 230-248.

[5] Priem, R., & Ashton, P. J. (2022). A european distributed ledger technology pilot regime for market infrastructures: finding a balance between innovation, investor protection and financial stability. Journal of Financial Regulation and Compliance, 30(3), 371-390.

[6] Jiang, T., Shen, Z., Jin, X., Zhang, R., Parisio, A., & Li, X., et al. (2023). Solution to coordination of transmission and distribution for renewable energy integration into power grids: an integrated flexibility market. CSEE Journal of Power and Energy Systems, 9(2), 444-458.

[7] Jiao, C., HongyanHao, Li, M., Fu, R., Liu, Y., & Lin, S., et al. (2023). Inter-provincial transaction model in two-level electricitymarket considering carbon emission and consumption responsibility weights. Energy Engineering, 120(10), 2393-2416.

[8] Petrillo, A., Murino, T., Piccirillo, G., Santini, S., & Caiazzo, B. (2023). An iot-based and cloud-assisted ai-driven monitoring platform for?smart manufacturing: design architecture and experimental validation. Journal of Manufacturing Technology Management, 34(4), 507-534.

[9] Rahman, A., Islam, M. J., Band, S. S., Muhammad, G., Hasan, K., & Tiwari, P. (2023). Towards a blockchain-sdn-based

secure architecture for cloud computing in smart industrial iot. Digital Communication and Networking: English Version, 9(2), 411-421.

[10] Yang, Y., Shi, Y., & Wang, T. (2022). Blockchain technology application maturity assessment model for digital government public service projects. International Journal of Crowd Science, 6(4), 184-194.

[11] Afzal, M., Li, J., Huang, Q., Umer, K., Ahmad, S. A., & Raza, A., et al. (2022). Role of blockchain technology in transactive energy market: a review. Sustainable Energy Technologies and Assessments, 36(4), 861-878.

[12] Chen, S., Ping, J., Yan, Z., Jinjin, L. I., & Huang, Z. (2022). Blockchain in energy systems: values, opportunities, and limitations. The Energy Frontier: The English version, 16(1), 10.

[13] Hasan, M., Alkhalifah, A., Islam, S., Babiker, N. B. M., Habib, A. A., & Aman, A., et al. (2022). Blockchain technology on smart grid, energy trading, and big data: security issues, challenges, and recommendations. Wireless Communications and Mobile Computing, 8(1), 198-211.

[14] YiDING, WeiSHEN, Hai-shengLI, Qiong-huiZHONG, Ming-yuTIAN, & JieLI. (2022). Blockchain trusted privacy service computing model for cnn. Acta Electronica Sinica, 50(06), 1399-1409.

[15] Li, J., Xing, Y., & Zhang, D. (2022). Planning method and principles of the cloud energy storage applied in the power grid based on charging and discharging load model for distributed energy storage devices. Processes, 10(2), 194.

[16] Drakatos, Koutrouli, & Tsalgatidou. (2022). Adrestus: secure, scalable blockchain technology in a decentralized ledger via zones. Blockchain Research, 3(4), 25.

[17] Sahbudin, M. A. B., Chaouch, C., Serrano, S., & Scarpa, M. L. (2021). Application-programming interface (api) for song recognition systems. Advances in Science Technology and Engineering Systems Journal, 6(2), 846-859.

[18] Alevizos, L., Ta, V. T., & Eiza, M. H. (2022). Augmenting zero trust architecture to endpoints using blockchain: a state-of-the-art review. Security and Privacy, 5(1), 191.

[19] Li, H., Gao, P., Zhan, Y., & Tan, M. (2022). Blockchain technology empowers telecom network operation. China Communications (English version), 19(1), 274-283.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: May 31, 2024
Accepted: Aug 2, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1222–1230, DOI 10.12694/scpe.v26i3.4118

LOAD DEMAND PREDICTION BASED ON IMPROVED ALGORITHM AND DEEP CONFIDENCE NETWORK

WEI XU[*], YI YU [†], YAQIN QIAN [‡], XU HUANG [§], MING ZHANG [¶], AND ZHONGPING SHEN [‖]

Abstract. To address the issue of inaccurate load forecasting amidst the advancing smart grid technology and the widespread integration of various demand-side resources like controllable loads, distributed energy sources, and energy storage, the author proposes a deep confidence network based on improved algorithms for load demand forecasting. Firstly, the VMD algorithm is used to decompose the load data into different intrinsic mode functions (IMFs), Then combine the DBN network to predict each IMF, Finally, overlay the prediction results of each part to obtain the prediction results of the VMD-DBN model. The experimental results indicate that: The PSO-DBN model has good prediction results and fast convergence speed in power load forecasting. The MAPE is 1.03%, and the RMSE is 9.35MW, which verifies that the method has good prediction accuracy. Compared to the single use of DBN method and the combination of Empirical Mode Decomposition (EMD) DBN method, the proposed method by the author has a significant improvement in prediction accuracy.

Key words: Short term load forecasting, Generalized demand side resources, Deep belief network, Load aggregator

1. Introduction. Short term load forecasting, as the foundation of daily power grid planning, is an indispensable and important part of safe power operation [1]. With the reform and development of electricity, improving the accuracy of short-term load forecasting has become the primary task of researchers in this field. It is of great significance for how to arrange scheduling plans, ensure reliable operation of power systems, and maximize economic benefits [2]. The past is a prophet of the future, and the basic principle of forecasting lies in fully learning the past. Therefore, the premise of load forecasting is also to predict the historical data of regional electricity loads. The amount of historical data required varies depending on the prediction period (the two are usually in a certain positive proportion), and combined with local politics, economy, meteorology, social events that affect electricity consumption, and other factors that can significantly affect electricity use, explore the necessary potential connections between changes in electricity loads and these influencing factors, in order to find a certain future development law and trend of electricity loads [3].

Short term load forecasting is of great significance for the optimal combination of units, economic dispatch, and optimal power flow of the dispatch department, especially for the current and future electricity market. Precise load forecasting enables strategic scheduling of power generation units, optimizing their efficiency and the economic viability of grid operations. This, in turn, fosters stability and security within the power grid. Within the smart grid framework, the integration of controllable loads, distributed energy sources, energy storage, respond to demand in a flexible and diverse manner, enhancing load transfer capabilities and expanding the time range for transfer [4]. In the electricity market environment, users adjust controllable loads, distributed power sources, and energy storage resources reasonably based on different price signals and incentive mechanisms with the goal of electricity economy, changing load characteristics and changing patterns. In short-term electricity market forecasting, it's crucial to factor in diverse demand-side resources to enhance accuracy. As the power grid expands and information technology advances, smart grid dispatch systems are becoming more adept at collecting vast amounts of data on loads and related factors, fueling exponential growth in data availability [5]. However, the load forecasting methods in the above literature are mostly shallow three-layer networks, which

[*] School of Common Courses, Wannan Medical College, Wuhu, 241000, China. (WeiXu966@163.com)

[†] Anhui ARN Group Co., LTD, Anqing, 246000, China. (Corresponding author, YiYu6375@126.com)

[‡] School of Common Courses, Wannan Medical College, Wuhu, 241000, China. (YaqinQian9@163.com)

[§] School of Common Courses, Wannan Medical College, Wuhu, 241000, China. (XuHuang3@126.com)

[¶] School of Common Courses, Wannan Medical College, Wuhu, 241000, China. (MingZhang16@163.com)

[‖] School of Common Courses, Wannan Medical College, Wuhu, 241000, China. (MingZhang16@163.com)

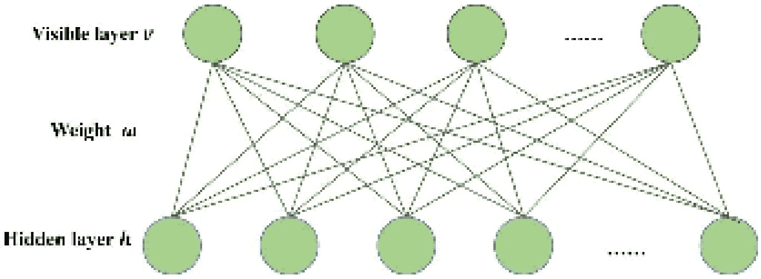


Fig. 3.1: RBM Structure

are difficult to handle the massive load dataset of the power grid today. Deep belief network (DBN) is an efficient deep learning algorithm composed of several stacked restricted Boltzmann machines (RBMs), which can be used for unsupervised learning and effectively process large power load data [6,7].

2. Literature Review. The flexible scheduling of controllable loads, distributed power sources, and energy storage, which are widely connected and participate in the electricity market, will inevitably affect the changes in electricity demand. Therefore, the author first constructs a contract based generalized demand side resource optimal scheduling model for three controllable resources: load curtailment (LC), load shift (LS), and energy storage (ES) [8]. This model aims to maximize the revenue of load aggregator (LA) and, under various constraints of the contract, solves the optimal scheduling strategy for generalized demand side resource participation in the electricity market based on real-time electricity prices. Yu, M. et al. introduced a short-term load forecasting model merging Fuzzy Exponential Weighting (FEW) with Improved Harmonic Search (IHS) algorithms. They validated the model's accuracy using fitness functions as evaluation criteria. Error analysis showed the model's effectiveness in predicting short-term electricity load data with strong stability and precision, offering valuable insights for implementing short-term forecasting in various industrial sectors[9]. To address data imbalance in ultra-short-term AC load forecasting, Tian, Z. et al. introduced a resistance-capacitance model featuring a two-phase parameter identification scheme[10]. In terms of short-term load forecasting in the power system, Jian, L. I. et al. tested a mutation model of RNN-LSTM. It effectively solves the problem of gradient explosion and disappearance caused by inputting a large amount of data in classical RNN [11]. Gao, W. et al. crafted a short-term load forecasting framework specifically tailored to accurately predict the cooling load of office buildings. They validated the framework's performance by assessing its ability to predict cooling loads, highlighting the importance of identifying key input features to enhance predictive accuracy[12].

Deep Belief Networks (DBN) are unsupervised learning models that converge faster and have higher prediction accuracy compared to traditional BP neural networks. A load forecasting algorithm based on DBN was proposed, combined with the fast optimization ability of particle swarm optimization, to achieve fast and accurate prediction of missing power data. Perform VMD on load data to obtain multiple sub sequences with distinct features, combine them with DBN prediction, and overlay each prediction result. By comparing the prediction results of a single DBN method and an EMD DNN prediction method, it was verified that the proposed method can explore the potential patterns of load data, reduce the computational scale, and reduce the generation of false IMFs, thereby improving prediction accuracy.

3. Method.

3.1. Restricted Boltzmann Machine.

Deep Belief Networks (DBNs) consist of multiple Restricted Boltzmann Machines (RBMs) stacked together. Each RBM in the model comprises interconnected hidden and visible layers arranged in a random combination. The connections between units in the hidden and visible layers are fully linked, facilitating robust unsupervised learning for data. The network structure is depicted in Figure 3.1, illustrating the absence of connections within individual hidden and visible layers.

If RBM uses v to represent the visible layer and h to represent the hidden layer, then the energy equation of the system is

$$E(v,h|\theta) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_j - \sum_{i=1}^{m}\sum_{j=1}^{n} v_i w_{ij} h_j \tag{3.1}$$

In the formula, v_i represents the state of visible layer unit i; h_j is the state of hidden layer unit j; a_i is the bias of visible layer unit i; b_j is the bias of hidden layer unit j; The number of all units in the visible layer m; n is the number of all units in the hidden layer; w_{ij} is the connection weight between units i in the visible layer and units j in the hidden layer; θ is the set of all parameters $\{a_i, b_i, w_{ij}\}$.

The joint probability distribution of a given state is

$$P(v,h|\theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)} \tag{3.2}$$

In the formula: $Z(\theta)$ is the partition function, represented as $Z(\theta) = \sum_v \sum_h e^{-E(v,h|\theta)}$.

Due to the independence between the units in the visible layer and the units in the hidden layer of RBM, its conditional probability distribution is:

$$P(v|\theta) = \frac{\sum_h e^{-E(v,h|\theta)}}{Z(\theta)} \tag{3.3}$$

$$P(v|\theta) = \frac{\sum_v e^{-E(v,h|\theta)}}{Z(\theta)} \tag{3.4}$$

RBM adopts unsupervised greedy training algorithm for parameter training, with the training objective of maximizing the logarithmic likelihood function of the model, a_i, b_j, and $lgP(v|\theta)$. By taking partial derivatives of the likelihood function and combining Gibbs sampling, the updated iteration formulas for the parameters a_i, b_i and w_{ij} can be obtained as follows:

$$\Delta a_i = \epsilon(< v_i >_{data} - < v_i >_{recon}) \tag{3.5}$$

$$\Delta b_i = \epsilon(< h_i >_{data} - < h_i >_{recon}) \tag{3.6}$$

$$\Delta w_{ij} = \epsilon(< v_i h_j >_{data} - < v_i h_j >_{recon}) \tag{3.7}$$

In the formula, $< \cdot >_{data}$ represents the mathematical expected value of the model distribution; represents the mathematical expected value of the distribution after further reconstruction of the model; $< \cdot >_{recon}$ ϵ is the learning rate [13,14].

3.2. Deep Confidence Network. DBN consists of multiple RBMs arranged in layers, with each RBM's hidden layer serving as the visible layer for the subsequent RBM in the stack. This hierarchical structure, depicted in Figure 3.2, facilitates the learning of increasingly abstract representations of data as it progresses through the network [15]. DBN adopts a greedy layer by layer training algorithm to complete the cognitive and generation process of the model from bottom to top, and then completes backpropagation training and weight fine-tuning from top to bottom through feedback learning of the classic BP neural network at the top.

3.3. Training models and learning algorithms. The time series prediction model adopted by the author is based on the DBN neural network algorithm. The model uses a DBN algorithm composed of multiple RBMs stacked together to perform forward unsupervised learning on the initial weights and thresholds. The greedy layer by layer training algorithm iteratively optimizes the various initial parameters of the training model, and then fine tunes the model parameters through classical feedback learning, so that the training results converge to the optimal. The model training process is shown in Figure 3.3.

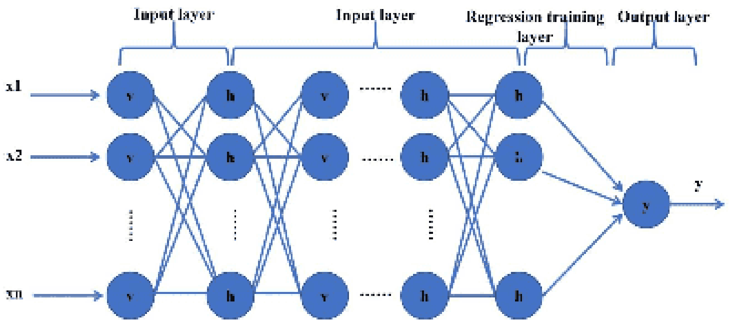


Fig. 3.2: DBN Model Structure

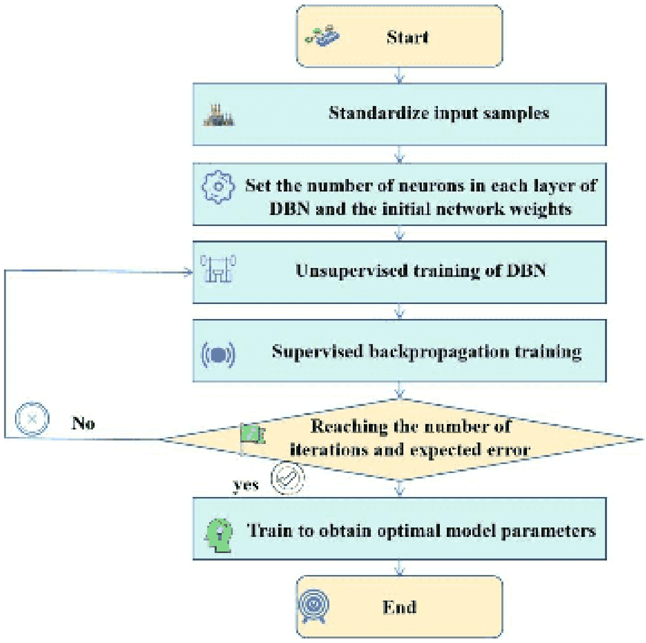

Start
Standardize input samples
Set the number of neurons in each layer of DBN and the initial network weights
Unsupervised training of DBN
Supervised backpropagation training
Reaching the number of iterations and expected error
No
yes
Train to obtain optimal model parameters
End

Fig. 3.3: DBN model training process

According to the temporal characteristics of load data, the experimental data is treated as a set of temporal data for model training. Assuming that the algorithm model has the $i^{\#}$-th input variable x_i^* and the i^*-th output variable y_i^*=x(t) at time t, among them, x(t) represents the timing value of the current time t, which is to use the value of that time as the output and the value of the previous $\hat{t}$ time periods as the input for prediction, that is

$$x_i^* = [x(t-\hat{t}), x(t-\hat{t}+1), \cdots, x(t-2), x(t-1)] \tag{3.8}$$

The specific training steps for the prediction model are as follows:

Step 1: Analyze and process the original power load data, and use the standard score formula to normalize the data to the [0,1] interval, as shown in equation 3.9. The normalized data can to some extent accelerate the convergence speed of the model and improve its accuracy.

$$x^* = \frac{x - \overline{x}}{\sigma} \tag{3.9}$$

In the formula, x^* represents the normalized data; x is the original experimental data; σ is a vector where each element represents the mean of the original data; is the standard deviation of the original experimental data.

Step 2: Use a DBN model stacked from multiple RBM models for data training, set the number of hidden layers n, and the learning rate of the BP neural network model during backpropagation training ϵ_{bp} and momentum factor α, and provide the number of DBN training times and the number of backpropagation algorithm training times. The condition for exiting the model iteration is to reach the maximum number of iterations or the expected error.

Step 3: To accomplish the unsupervised learning process in the DBN model, a greedy layer-by-layer training algorithm is employed. As the efficiency of Gibbs sampling diminishes with more sampling steps, the author opts for the Contrastive Divergence (CD) algorithm, introduced by Hinton, for swift parameter estimation.

The CD algorithm can obtain sufficiently good training model parameters through a one-step sampling method [16]. Based on the symmetric structure and independence of the model, the activation probability $P(h|v_0)$ and the initial state h0 of the hidden layer are obtained by using the initial state v_0 of the visible layer. After a step of Gibbs sampling, v_1 and h_1 can be obtained based on the initial state of the model. The specific sampling process is as follows.

The sigmoid function is used as the excitation function between the neurons in the hidden layer of the model to standardize the data. The formula for the processed function is

$$\hat{\sigma}(\dot{y}) = \frac{1}{1+e^{-y}} \tag{3.10}$$

In the formula, $\dot{y}$ represents the data to be subjected to sigmoid standardization processing.

From this, we can obtain the activation probability distribution of the visible layer and the hidden layer when they are turned on:

$$P(h_j = 1|v, \theta) = \hat{\sigma}(b_j + \sum_i v_i w_{ij}) \tag{3.11}$$

$$P(v_j = 1|h, \theta) = \hat{\sigma}(a_j + \sum_i w_{ij} h_j) \tag{3.12}$$

Finally, the various training parameters of the model can be updated according to equations 3.5-3.7 and 3.13-3.15 as follows:

$$w_{ij}^{k+1} = w_{ij}^k + \Delta w_{ij} \tag{3.13}$$

$$a_{ij}^{k+1} = a_i^k + \Delta a_i \tag{3.14}$$

$$b_{ij}^{k+1} = b_j^k + \Delta b_j \tag{3.15}$$

In the formula, k represents the number of iterations.

3.4. Prediction Model Based on Particle Swarm Optimization. The Particle Swarm Optimization (PSO) algorithm finds wide application in power data prediction tasks. It enhances the convergence speed and accuracy of training models by ensuring they reach global optimal solutions. PSO optimizes and updates model parameters, simulating the collective foraging behavior of birds. In this analogy, each particle in the PSO model represents an individual bird, navigating the solution space by adjusting its position and movement speed to find local optimal solutions. The particle swarm obtains a global optimal solution by sharing information between each particle. Assuming that the position of the k-th iteration of the particle swarm is $Z_i^k = (Z_{b1}^k, Z_{b2}^k, \cdots, Z_{bd}^k)$, the speed of movement is $U_b^k = (U_{b1}^k, U_{b2}^k, \cdots, U_{bd}^k)$, the optimal position of each particle is represented as $P_b^k = (P_{g1}^k, P_{g2}^k, \cdots, P_{gd}^k)$, and the optimal position of the entire particle swarm is represented as $P_b^k = (P_{b1}, P_{b2}, \cdots, P_{bd})$, the update formula for the position and movement speed of each particle can be obtained as follows:

$$Z_{bd}^{k+1} = Z_{bd}^k + U_{bd}^{k+1} \tag{3.16}$$

$$U_{bd}^{k+1} = \omega U_{bd}^k + c_1 r_1 (P_{bd} - Z_{bd}^k) + c_2 r_2 (P_{gd} - Z_{bd}^k) \tag{3.17}$$

Building on the characteristics of the PSO algorithm, the author introduces the PSO-DBN model. This model optimizes the parameters further following the unsupervised training of DBN. The formula involves model weights, learning factor constants (c1 and c2), random numbers (r1 and r2) within [0,1], the number of particles (d), and the number of iterations (k). Through this approach, the PSO-DBN model refines the parameter settings, enhancing the overall performance of the system. The initial particle swarm position of the PSO model is trained using DBN parameters, and then PSO iteratively optimizes the regression training layer of the model. The training parameters are the connection weight w1 and bias of the first layer, respectively 1, as well as the connection weight w_2 and bias θ_2 of the second layer, the update formula for the training parameters is:

$$w_1^{k+1} = \begin{bmatrix} Z_{b1}^k & \cdots & Z_{bs_1}^k \\ \vdots & \ddots & \vdots \\ Z_{b(s_1 s_2 + s_1 - 1)}^k & \cdots & Z_{b(s_1 s_2 + 2s_1)}^k \end{bmatrix} \tag{3.18}$$

$$\theta_1^{k+1} = \left[Z_{b(s_1 s_2 + 2s_1 + 1)}^k \cdots Z_{b(s_1 s_2 + 3s_1)}^k \right] \tag{3.19}$$

$$w_2^{k+1} = \left[Z_{b(s_1 s_2 + 2s_1 + 1)}^k \cdots Z_{b(s_1 s_2 + 3s_1)}^k \right] \tag{3.20}$$

$$\theta_2^{k+1} = Z_{b(s_1 s_2 + 3s_1 + 1)}^k \tag{3.21}$$

In the formula: k is the number of iterations; s_1 is the number of units in the first layer of DBN; s_2 is the number of units in the second layer of DBN [17,18]. The PSO-DBN model training flow is shown in Figure 3.4.

4. Results and Discussion. Modeling and analyzing the distribution network load data obtained from the project, mainly focusing on model training and prediction of historical values of distribution network load detection values. Due to the differences in data between different equipment points in substations, in order to make the prediction model universal for different point data in substations, the measurement data of a point in the database is selected as the experimental object, and the experimental data is normalized to improve the training speed and accuracy of the training model. In order to better evaluate the predictive accuracy of the prediction model in power load forecasting, two indicators, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), were used to analyze the accuracy of the experimental results[19-20]. The specific indicators are calculated as follows:

$$\hat{E}_{RMSE} = \sqrt{\frac{\sum (x_e - x_e')^2}{n_e}} \tag{4.1}$$

$$\hat{E}_{RMSE} = \frac{1}{n_e} \sum \left| \frac{x_e - x_e'}{x_e} \right| \tag{4.2}$$

In the formula, x_e represents the real data; x_e' is the predicted data; n_e is the total number of data.

The simulation comparison results between two prediction models proposed by the author and the BP neural network model are shown in Table 4.1.

Figure 4.1 shows the convergence curves of three models. Based on the simulation results in Table 1 and the time complexity of the three models, it can be seen that the PSO-DBN model has good prediction results and fast convergence speed in power load forecasting. The MAPE is 1.03%, and the RMSE is 9.35MW, which verifies that the method has good prediction accuracy.

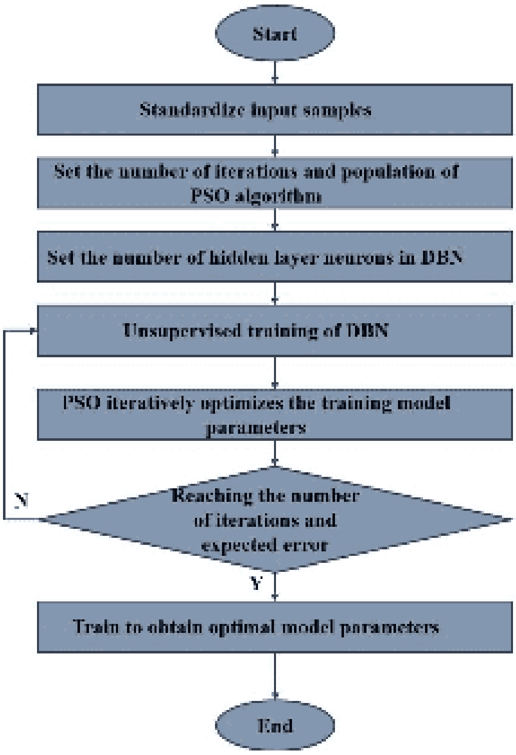

Start
Standardize input samples
Set the number of iterations and population of PSO algorithm
Set the number of hidden layer neurons in DBN
Unsupervised training of DBN
PSO iteratively optimizes the training model parameters
Reaching the number of iterations and expected error
N
Y
Train to obtain optimal model parameters
End

Fig. 3.4: PSO-DBN model training process

Table 4.1: Simulation Results

training model	RMSE	MAPE
BP	2.2338	0.0052
DBN	1.3063	0.0031
PSO-DBN	1.2763	0.0028

5. Conclusion. The author proposes a load demand prediction based on an improved algorithm using deep confidence networks. The participation of a large number of generalized demand side resources in the electricity market has higher requirements for short-term load prediction accuracy. At the same time, the massive dataset generated by the smart grid dispatch system provides a data foundation for the use of deep learning. Hence, the author initially integrates generalized demand side resources into market operations via load aggregation merchants, establishing a contract-based scheduling model for generalized demand side resources to derive optimal scheduling strategies. Then, the optimal scheduling scheme for generalized demand side resources is used as input for a load prediction model. In this paper, a DBN short term load forecasting model, which includes generalized demand side resources, is developed and compared with the BP neural network and the DBN model. The empirical results demonstrate the efficacy of the demand response resource scheduling model, centered on electricity price contracts, in maximizing revenue for load aggregators. It effectively adapts to real-time market electricity prices, determining optimal participation times for various resources. Moreover, integrating the optimal scheduling plan for generalized demand side resources into the prediction model proves advantageous, enhancing prediction accuracy and reducing errors.

Acknowledgement.

1. 2020 Anhui Province Quality Engineering Project: Exploration and Practice of Innovation and En-

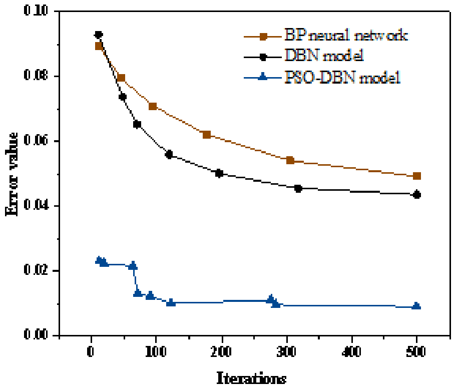


Fig. 4.1: Convergence curves of three models

trepreneurship Training of regional medical college students from the perspective of Yangtze River Delta Integration Development (2020jyxm2076).

2. 2022 Wannan Medical College Teaching Quality and Teaching Reform Project Research topic: Medical College moral education and Innovation and entrepreneurship education integration practice path and evaluation reform research (2022jyxm21).

REFERENCES

[1] Sun, F., Huo, Y., Fu, L., Liu, H., Wang, X., & Ma, Y. (2023). Load-forecasting method for ies based on lstm and dynamic similar days with multi-features. Global Energy Internet: English, 6(3), 285-296.

[2] Zhang, R., Yu, M., & Zhang, C. (2022). A similar day based short term load forecasting method using wavelet transform and lstm. IEEJ Transactions on Electrical and Electronic Engineering, 17(4), 506-513.

[3] Hao, L., Linghua, Z., Cheng, T., & Chenyang, Z. (2023). Short-term load forecasting model based on gated recurrent unit and multi-head attention. Journal of China University of Posts and Telecommunications: English Edition, 30(3), 25-31.

[4] Guo-Feng Fan, Li, Y., Xin-Yan Zhang, Yi-Hsuan Yeh, & Wei-Chiang Hong. (2023). Short-term load forecasting based on a generalized regression neural network optimized by an improved sparrow search algorithm using the empirical wavelet decomposition method. Energy Science And Engineering, 11(7), 2444-2468.

[5] Xiao, L. I., & Xianling, L. U. (2022). Ethod for forecasting short-term power load based on dual-stage attention mechanism and gated recurrent unit network. Computer Engineering, 48(2), 291-296.

[6] Zhang, R., Yu, M., & Zhang, C. (2022). A similar day based short term load forecasting method using wavelet transform and lstm. IEEJ Transactions on Electrical and Electronic Engineering, 17(4), 506-513.

[7] Xin, L., Haidong, S., Hongkai, J., & Jiawei, X. (2022). Modified gaussian convolutional deep belief network and infrared thermal imaging for intelligent fault diagnosis of rotor-bearing system under time-varying speeds:. Structural Health Monitoring, 21(2), 339-353.

[8] Su, S., Hu, G., Li, X., Li, X., & Xiong, W. (2023). Electricity-carbon interactive optimal dispatch of multi-virtual power plant considering integrated demand response. Energy Engineering (English), 120(10), 2343-2368.

[9] Yu, M., Zhu, J., & Yang, L. (2023). Short-term load prediction model combining few and ihs algorithm. Archives of Electrical Engineering, 9(2), 433-443.

[10] Tian, Z., Lin, X., Lu, Y., Song, W., & Niu, J. (2023). Imbalanced data-oriented model learning method for ultra-short-term air conditioning load prediction. Energy and buildings, 11(5), 695.

[11] Jian, L. I., Peng, Y., Cheng, Z., Yuwei, L. I., Fan, J., & Chen, J. (2022). Prediction and analysis on short-term load of power system based on lstm. Meteorological and Environmental Studies: English Version (004), 013.

[12] Gao, W., Huang, X., Lin, M., Jia, J., & Tian, Z. (2022). Short-term cooling load prediction for office buildings based on feature selection scheme and stacking ensemble model. Engineering computations: International journal for computer-aided engineering and software(5), 39.

[13] Zhang, R., Yu, M., & Zhang, C. (2022). A similar day based short term load forecasting method using wavelet transform and lstm. IEEJ Transactions on Electrical and Electronic Engineering, 17(4), 506-513.

[14] Dong, S., & Xia, Y. (2023). Network traffic identification in packet sampling environment. Digital Communication and Networking: English Version, 9(4), 957-970.

[15] Gowri, S., Subhashini, R., Mathivanan, G., Jabez, J., Vigneshwari, S., & Vimali, J. S. (2024). A descriptive framework for information retrieval using crawler based clustering and effective search algorithm. International Journal of Information Technology & Decision Making, 23(02), 993-1016.

[16] Kayij, E. N., Joél Lema Makubikua, & Busili, J. D. K. (2023). New hybrid algorithm based on bicriterionant for solving multiobjective green vehicle routing problem. American Journal of Operations Research, 13(3), 33-52.

[17] Babu, R. G., Amudha, V., Chellaswamy, C., & Kumar, K. S. (2022). Retracted article: lot based residential energy management system for demand side response through load transfer with various types of domestic appliances. Building Simulation (English), 15(9), 1.

[18] Chen, G., Huang, W. X., Ronch, A. D., Pecora, R., Kim, D., & Liu, Y., et al. (2023). Bp neural network-kalman filter fusion method for unmanned aerial vehicle target tracking:. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 237(18), 4203-4212.

[19] Yang, J., Zhang, L., Liu, G., Gao, Q., & Qian, L. (2022). Sintered silicon carbide grinding surface roughness prediction based on deep learning and neural network. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 44(7), 1-13.

[20] Ma, G., Wang, Z., Liu, W., Fang, J., Zhang, Y., & Ding, H., et al. (2023). Estimating the state of health for lithium-ion batteries:a particle swarm optimization-assisted deep domain adaptation approach. Journal of Automation: English Edition, 10(7), 1530-1543.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 10, 2024
Accepted: Aug 2, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1231–1241, DOI 10.12694/scpe.v26i3.4122

INTERPRETABLE AI ARCHITECTURE OF MACHINE LEARNING ALGORITHM FOR INTELLIGENT VIDEO SURVEILLANCE BASED ON FOG AND EDGE COMPUTING

JIE ZHANG[*], WEIPING SONG[†], WENKUI SHE[‡], HUANHUAN LI[§], FEIHU HUANG[¶], AND FAN YANG[||]

Abstract. In order to address the application scenarios and technical requirements of video monitoring, the centralized data processing model represented by cloud computing has a large cost in terms of resource requirements, relies excessively on the network bandwidth of the cloud computing center, and is difficult to meet the needs of video processing in real-time and other aspects, the author proposes an interpretable AI architecture of intelligent video monitoring machine learning algorithm based on fog and edge computing. The author proposes a edge computing model suitable for video monitoring scenarios. From the three main resources of computing, network bandwidth and storage as the entry point, the system architecture is designed. After using the computing power of edge nodes to complete video preprocessing, the Docker container platform is built, and hierarchical scheduling strategy is adopted to reduce network congestion. The results showed that compared to video surveillance with added motion detection function, analyzing data from week 1 to week 3 compared to video surveillance without added motion detection function, the number of video files stored was about 1176 records per week, saving about 41.56% of storage space. This model can effectively reduce the computational, storage, and network transmission costs in video surveillance scenarios.

Key words: Edge computing model, Intelligent video surveillance, Docker container, scheduling strategy

1. Introduction.

Fog computing is a concept proposed by Cisco in the United States in 2011. It is a distributed computing oriented towards the Internet of Things, aimed at supplementing and improving cloud computing [1]. Compared to "cloud computing," fog computing can more vividly describe its style of being close to the ground and vast permeability, closer to people's surroundings, and more down-to-earth. The original intention of fog computing design is to deploy a large number of devices with average computing power and low cost, thereby extending large-scale networked computing power and data analysis capabilities to the edge of the network, allowing users to receive fast network service feedback and data feedback. Edge computing is one of the important technologies to realize fog architecture. The European Telecommunications Standards Institute (ETSI) first gave an explanation of edge computing at mobile terminals and provided a reference architecture for edge computing in 2014 [2]. By 2025, Gartner forecasts that the majority — 75% — of enterprise-generated data will be handled outside of conventional cloud platforms. Beginning March 6, 2020, AT&T entered into a partnership with Alphabet's Google Cloud to leverage 5G edge computing capabilities. This collaboration aims to enhance customer experiences by deploying applications nearer to end-users, thereby bolstering speed and security [3].

As one of the terminal devices with the greatest demand for network bandwidth, storage and computing capacity at this stage, video monitoring is a typical scenario of the above edge computing applications. Traditional video monitoring systems generate video stream data, which is mostly transmitted to the business center in real time through IP networks. Edge computing has obvious advantages not only in solving the bottleneck of processing capacity in traditional video monitoring systems, improving system capacity and quality of service, but also in emerging intelligent video monitoring systems. On the one hand, using the idea of edge computing can process data near video monitoring equipment, reduce data sending to the cloud, reduce bandwidth, cloud

[*] Aostar Information Technologies Co., Ltd., Chengdu, 610041, China.(Corresponding author's email:JieZhang19@126.com)

[†] Aostar Information Technologies Co., Ltd., Chengdu, 610041, China.(WeipingSong7@163.com)

[‡] Aostar Information Technologies Co., Ltd., Chengdu, 610041, China.(WenkuiShe@126.com)

[§] Aostar Information Technologies Co., Ltd., Chengdu, 610041, China.(HuanhuanLi9@163.com)

[¶] Aostar Information Technologies Co., Ltd., Chengdu, 610041, China. College of Computer Science, Sichuan University, Chengdu, 610065, China.(FeihuHuang6@126.com)

[||] Aostar Information Technologies Co., Ltd., Chengdu, 610041, China.(FanYang966@163.com)

computing and storage pressure, and reduce latency and improve QoS for online processing scenarios [4,5]. On the other hand, AI technologies such as data desensitization and federated learning can be combined to improve the system's capability level. These are the reasons for applying edge computing in the field of intelligent video surveillance.

2. Literature Review. The idea of edge computing is to propose content distribution network technology in order to solve the problems of small network bandwidth, large and uneven user access and other problems existing in its own business, deploy cache servers close to users according to geographical location, redirect user requests to cache servers, reduce the pressure on central servers, reduce network latency, and improve service quality [6]. Yu, J. et al. developed a BP neural network model that is enhanced by a genetic algorithm, with edge computing serving as the central component. They validated the effectiveness of this model through case studies. Their practical analysis demonstrates that the predictions generated by this model exhibit higher accuracy and alignment with real-world power grid scenarios. As a result, this model holds promise for practical application in various contexts [7]. Song, W. et al. introduced an innovative approach termed Intelligent Standing Human Detection (ISHD). This method relies on an enhanced Single Shot Multi-Box Detector (SSD) to identify standing human poses within video surveillance frames, particularly in exam stage environments[8]. Isaac Martín de Diego. et al. devised a novel method aimed at enhancing existing intelligent video surveillance systems by imbuing them with environmental sensitivity. This method leverages expert guidance to tailor the system's behavior to specific scenarios, incorporating data for generating alerts and contextual understanding. Encouraging experimental outcomes underscore the promise of this approach, positioning it as a foundation for future system improvements [9]. Khosravi, M. R. et al. investigated various facets of multimedia computing within the realm of Video Synthetic Aperture Radar (SAR), an emerging real-time remote sensing and surveillance radar imaging mode [10].

The author proposes to deploy the video capture framework and container application based on edge computing in the smart campus system. In the video surveillance system based on the edge computing model, edge nodes are used to deploy motion detection algorithms to preprocess the video streams collected by edge devices, such as moving object detection, so as to reduce redundant information and reduce the system's demand for storage and transmission. Deploying a container platform brings better scheduling features, solves the resource scheduling problem of edge computing nodes, and improves the overall performance of the cloud computing center.

3. Research Methods.

3.1. Requirements for motion detection function. In monitoring scenarios, there are often a large number of cameras. If abnormal situations need to be detected in a timely manner, a large number of manual personnel are usually arranged to conduct real-time 7$*$ 24 hours of video inspection. Meanwhile, due to the persistent nature of video information flow, the transmission of video information will also bring huge network loads; Over time, the video data generated by monitoring systems will also bring enormous storage pressure [11]. If these massive video data are directly uploaded to cloud computing centers, on the one hand, the processing of video information requires a large amount of computing resources, and on the other hand, the transmission and storage of data will also face enormous pressure.

The main purpose of storing video information in monitoring scenarios is to record changes and suspicious information in the scene. If appropriate techniques or methods are not used, the monitoring system will occupy a large network bandwidth, transmit long-term records of unchanged monitoring scenes, have low effective information content, and video information will lose its storage significance [12]. Therefore, in intelligent video surveillance systems, motion detection technology is the most basic and important technology. This technology detects moving targets in video data streams through appropriate algorithms, replacing manual recognition work. By setting certain parameters, the motion characteristics or position information of moving targets in the monitoring scene can be discovered, achieving automatic alarm or determining whether video information meets the requirements of storage or transmission backup, which can effectively save storage space and reduce network transmission pressure. Among various motion detection algorithms, optical flow method, background difference method, and inter frame difference method are the most common.

3.1.1. Optical flow method. The optical flow technique correlates a two-dimensional velocity field with grayscale images, incorporating constraint equations to derive the fundamental algorithm for optical flow computation. While this method is relatively straightforward and simple to deploy, fluctuations in lighting conditions or occluded objects can distort the optical flow field, thereby amplifying the computational demands of the algorithm. Consequently, these limitations pose challenges when applying optical flow in real-time scenarios [13,14].

3.1.2. Background difference method. The background difference method selects a specific image as the background frame, and then performs a difference operation between the current video frame or image to be judged and the background frame, in order to further determine whether there is a moving target, the processing process of the background difference algorithm is as follows:

(1) Select the image without moving objects entering the monitoring screen as the background frame, defined as $background(x,y)$;

(2) Select the current frame that requires comparison and judgment, defined as $frame_k(x,y)$;

(3) Set the threshold to T, perform a difference operation between the current frame $frame_k(x,y)$ and the background frame $background(x,y)$, compare the difference result with the threshold T, and binarize to obtain the moving target. If it is greater than the threshold T, it is judged that there is a moving target. If it is less than or equal to the threshold T, it is judged that there is no moving target. The formal calculation formula is expressed as follows 3.1:

$$detext(x,y) = \begin{cases} 1 & \text{if } |frame_k(x,y) - background(x,y)| > T \\ 0 & \text{others.} \end{cases} \tag{3.1}$$

$detext(x,y)$ is the binary image obtained by differential operation and binarization between the current frame and the background frame. Only when $detext(x,y) = 1$ is used, it indicates the detection of a moving target.

The background difference method only requires differential detection of one frame, which is fast and accurate. However, the background subtraction algorithm largely relies on the reliability of the background frame $background(x,y)$. If there are changes in lighting, shadows, etc., it is necessary to continuously adjust the background frame to adapt to changes in the environment. Therefore, the background subtraction algorithm is more suitable for fixed cameras [15].

3.1.3. Two frame difference method. The two frame difference method, also known as the inter frame difference method, has a similar algorithm design approach to background difference. It adopts an improved approach to select adjacent two frames of images, grayscale the current frame $frame_k(x,y)$ and the previous frame $frame_{k-1}(x,y)$, and perform differential operation. The inter frame difference method is not affected by slow light changes, and the algorithm is simple and easy to implement. The formal calculation formula is expressed as follows 3.2:

$$detext(x,y) = \begin{cases} 1, & \text{if } |frame_k(x,y) - frame_{k-1}(x,y)| > T \\ 0, & \text{others.} \end{cases} \tag{3.2}$$

But only the parts that change before and after the two frames can be detected, and the overlapping parts cannot be detected, which can easily lead to problems such as blurred and incomplete edges. When the object moves slowly, misjudgment or hollow phenomena may occur [16].

3.1.4. Three frame difference method. On the basis of the two frame difference method, researchers have proposed the three frame difference method. The basic idea is to extract consecutive three frames of images $frame_{k-1}(x,y)$, $frame_k(x,y)$, and $frame_{k+1}(x,y)$ The algorithm flow is shown in Figure 3.1.

The algorithm calculation process is as follows:

(1) Perform inter frame difference operation between frame k-1 and frame k according to formula 3.2 to obtain $detext_1(x,y)$;

(2) Perform inter frame difference operation between frame k and frame k+1 according to formula 3.2 to obtain $detect_2(x,y)$;

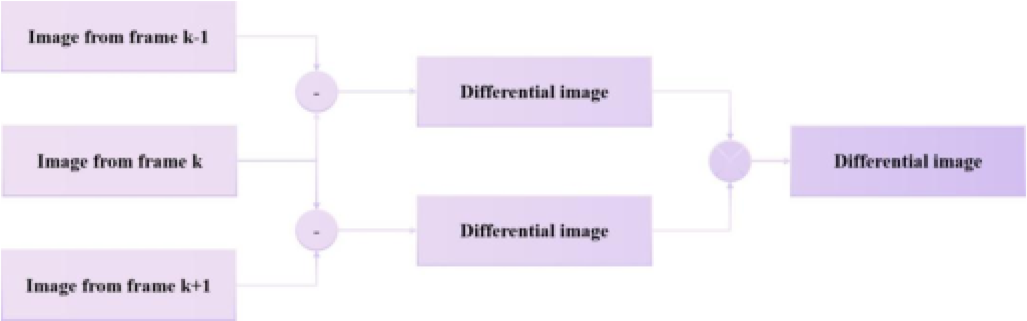


Fig. 3.1: Three frame differential algorithm flowchart

(3) The calculation results of $detect_1(x,y)$ and $detect_2(x,y)$ are combined, and the formal calculation formula is expressed as follows 3.3:

$$DETE(x,y) = detect_1(x,y) \bigotimes detect_1(x,y) = \begin{cases} 1 & detect_1(x,y)detect_1(x,y) \neq 0 \\ 0, & \text{others.} \end{cases} \tag{3.3}$$

Among them, $DETE(x,y)$ is the result of logical AND operation.

Similar to the two frame difference method, the three frame difference method still has the phenomenon of holes in the process of detecting moving targets. However, the three frame difference method can locate the position of the moving target in the monitoring screen, improving the accuracy of motion detection. The detection results are more accurate than the two frame difference method [17].

Based on the analysis of the above motion detection algorithms, in the author's research process, the three frame difference method can be used to implement the motion detection module, build a video monitoring system for the campus network, reduce the difficulty of system construction, achieve selective storage of video information, filter out effective video frames, reduce redundant information in the monitoring video, and achieve the goal of improving the effectiveness of video information. At the same time, it reduces the storage space and network bandwidth expenses of the massive video information obtained by a large number of video capture devices, achieving the goal of saving construction costs.

3.2. Analysis of storage and network bandwidth requirements.

3.2.1. Camera stream analysis. In order to meet the construction requirements of smart campuses, high-definition cameras are mainly used in the monitoring system. The specifications of cameras vary, and the amount of data generated also varies. Taking HD digital cameras as an example, calculated based on a 2048 Kbps stream, each camera generates approximately 900 M of video data per hour and approximately 21 GB of new video data per day.

3.2.2. Analysis of Video Data Transmission and Storage. Taking the teaching building as an example, the floor structure is in a zigzag shape, with corridors above and below, and 5 classrooms distributed on each side. There are 10 classrooms on each floor, with a total of 4 corridors up, down, left, and right. The construction cost of the comprehensive monitoring system includes installing one FHD digital camera at each end of the four corridors, and installing two HD digital cameras in each classroom. A total of 28 cameras of two specifications are installed on each floor. When the system is running normally, each camera in the intelligent video monitoring system will generate two real-time data streams, which are used to monitor the video data stream transmitted in real-time and the video data file transmitted to the storage data stream of the cloud computing storage center. When an emergency occurs, the staff of the monitoring center need to view the video surveillance content in real time, and the storage center of cloud computing will also store video information in real time. The network bandwidth of the monitoring system will reach its maximum demand.

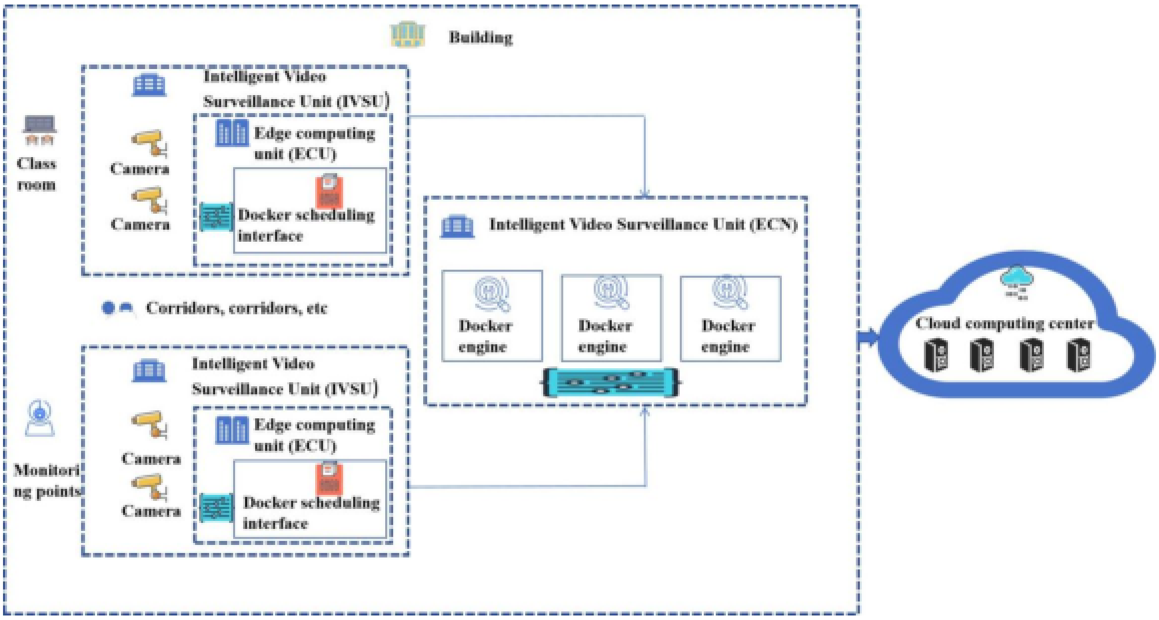

Building
Class room
Intelligent Video Surveillance Unit (IVSU)
Camera
Camera
Edge computing unit (ECU)
Docker scheduling interface
Corridors, corridors, etc
Intelligent Video Surveillance Unit (IVSU)
Monitoring points
Camera
Camera
Edge computing unit (ECU)
Docker scheduling interface
Intelligent Video Surveillance Unit (ECN)
Docker engine
Docker engine
Docker engine
Cloud computing center

Fig. 3.2: Architecture of intelligent video monitoring system based on edge computing model

When an emergency situation occurs, the maximum real-time video data stream and maximum storage data stream generated on each floor are: 4 Mbps ∗ 8 channels+2 Mbps ∗ 20 channels=72 Mbps [18]. Therefore, calculated by floor, the maximum network bandwidth requirement is a total of 144 Mbps, and the maximum storage write rate requirement for cloud computing centers is 72 Mbps. Deploying according to the same construction specifications, if an emergency occurs in the entire building, calculated based on 6 floors, the maximum transmission bandwidth needs to be 864 Mbps, and the maximum storage write rate needs to be 432 Mbps.

3.3. Management Requirements Analysis. Intelligent video surveillance systems generally include video capture, image preprocessing, motion object detection, motion object tracking, motion object classification, behavior description and understanding, and alarm processing modules. The corresponding management functions are also designed around these modules to improve the accuracy and response speed of the system, enhance the overall reliability of the system, and provide convenient management functions for managers. Compared with the traditional IVSS, the video monitoring system based on the edge computing model, in addition to the functions involved above, uses the edge computing model to pre process the video information at the edge nodes in a decentralized manner and store initial data. Therefore, the design adopts container technology to build a containerized resource scheduling platform, adopts appropriate strategies to achieve scheduling control of network bandwidth resources, uploads video files to cloud computing centers, realizes backup storage of video files, reduces network load and storage space requirements, improves resource utilization, and achieves the goal of reducing system construction costs [19].

3.4. Architecture Design. The architecture diagram is designed based on the structure of the teaching building. The intelligent video monitoring system based on the edge computing model is planned to have the following four roles: Edge computing Unit (ECU), Intelligent Video Surveillance Unit (IVSU), edge computing Node (ECN), Cloud Computing DataCenter (CCDC). The system architecture is shown in Figure 3.2.

(1) ECU. Edge computing unit, with certain computing power, can preprocess the video information collected by the camera, and provide file storage and network transmission functions. In the subsequent model validation, Raspberry Pi Zero W single board computer was used, with Linux installed at the bottom to provide Docker scheduling interface for the future.

(2) IVSU. Intelligent video monitoring unit, based on the ECU module, installs motionEyeOS and CSI cameras to achieve intelligent video monitoring unit, deploys video acquisition points, such as classrooms, corridors, training rooms, libraries, etc.

(3) ECN. The edge computing node has high computing power and can provide large storage space for temporary or long-term storage of video data generated by IVSU. It can select servers or other general-purpose computers as the hardware support environment according to the number of cameras, which is easy to deploy a container platform to support subsequent resource scheduling and model verification.

(4) CCDC. Cloud computing data center, deploying a large number of servers and storage hardware, using platforms or tools such as KVM and VMware, to build a cloud computing resource management center, and store massive video information for intelligent video monitoring systems.

3.5. Video capture framework design. When designing a video capture framework, Raspberry Pi is used as the hardware infrastructure for the framework while ensuring availability, in order to reduce construction costs. Deploy containers based on open-source platforms and provide support for scheduling strategies in a decentralized environment.

3.5.1. Open Hardware Architecture. Raspberry Pi is an open, easily scalable small single board computer with low power consumption, customizable on demand, providing all expected features or capabilities. Widely used in real-time image and video processing, as well as various IoT based applications. The video capture framework adopts edge nodes with Raspberry Pi Zero W as the core component, and loads CSI cameras to achieve the video capture function of the monitoring system.

Raspberry Pi Zero W single board computer, compact design, low power consumption, powered by Micro USB interface, low cost for video capture. This single board computer uses BCM2835 as the SoC, integrating various functions from general-purpose computers. In the calculation module, ARM1176JZF-S is used, providing 700MHz of computing power; In the video module, Broadcom VideoCore IV technology is adopted, which can achieve 1080p H.264 video encoding or decoding at 30 frames per second, while providing miniHDMI output function; In terms of network connection, WiFi and Bluetooth modules are provided, supporting 802.11n connection; In terms of other interfaces, it provides a mini USB On the Go interface, a Micro SD card interface, and a 40pin GPIO interface. For cameras, it provides a CSI interface that can adapt to Raspberry Pi Camera Module V2 cameras and capture high-definition videos up to 8 million pixels.

3.5.2. Open source video capture system. MotionEyeOS is an embedded operating system that uses the BuildRoot tool to complete cross compilation. It is suitable for deployment on single board computers and provides the implementation of a complete video surveillance system. The front-end of the video surveillance system is a motionEye program written in Python, providing web access functionality; The backend adopts a highly configurable motion program, which can view video streams in real time, and also achieve functions such as facial recognition, dynamic monitoring, camera direct recording, recording activity images, and creating dynamic video files. Improve motionEyeOS, simplify the process, use three frame difference algorithm, and achieve motion detection function [20].

3.6. Design of containerized resource scheduling scheme. In the design process of scheduling schemes, load is the main factor affecting the demand for application resources. Combining with the practical application of intelligent video surveillance systems, the system bottleneck mainly focuses on the network and disk I/O. Therefore, when designing a scheduling plan, the characteristics of the Docker container engine should be utilized first to provide information support for containerized scheduling plans by periodically collecting resource load information such as CPU, memory, disk I/O, and network bandwidth. In terms of scheduling mode, it fully reflects the decentralized processing characteristics of the edge computing model. For the architecture shown in Figure 3.1, a two-level scheduling mode is adopted, that is, the cloud computing data center (CCDC) schedules the edge computing node (ECN), and the edge computing node (ECN) schedules the edge computing unit (ECU). When executing scheduling tasks, the initiator is the Active Scheduling Object (ASO) and the other is the Passive Scheduling Object (PSO).

$$Aso_i = [A_{cpu} A_{mem} A_{net} A_r A_w A_{tasked}]^T \tag{3.4}$$

In equation 3.4, Aso_i represents the i-th active scheduling object (ASO) in the corresponding CCDC or ECN. A_{cpu} represents the remaining CPU resources in ASO, A_{mem} represents the remaining memory in ASO, and A_{net} represents the remaining network bandwidth in ASO. The above three resources are calculated as percentages, with a value range of $(0, 100)$; A_r represents the read status of disk I/O operations in ASO; A_w represents the write status of disk I/O operations in ASO;A_{tasked} indicates whether a scheduling task is being executed by a higher level in the current ASO. The result is a logical value set to (Ture | False), where Ture indicates that the scheduling task is currently being executed and False indicates that it has not been executed.

$$P_{so_j} = [P_{cpu}P_{mem}P_{net}P_rP_wP_{tasked}P_{resp}P_{ltime}P_{size}]^T \tag{3.5}$$

In equation 3.5, P_{so_j} represents the jth passive scheduling object (PSO) corresponding to the ECN or ECU. P_{cpu} represents the remaining CPU resources in PSO, P_{mem} represents the remaining memory resources in PSO, and P_{net} represents the remaining network bandwidth in PSO. Similar to formula 3.4, the above three resources are calculated as percentages, with a value range of (0,100); P_r represents the read status of disk I/O operations in PSO, while P_w represents the write status of disk I/O operations in PSO; P_{tasked} indicates whether the PSO is being scheduled by the previous level, and the result is a logical value set to (Ture | False). Ture indicates that the scheduled task is currently being executed, and False indicates that the scheduled task has not been executed; P_{resp} represents the network state between PSO and ASO; P_{ltime} represents the time when PSO was last successfully executed for scheduling tasks, and P_{size} represents the file size that needs to be scheduled for processing.

$$Task_{ij} = [T_{stime}T_{pri}T_{exectime}]^T \tag{3.6}$$

In equation 3.6, $Task_{ij}$ represents the execution parameters of the active scheduling object Aso_i initiating scheduling tasks to the passive scheduling object P_{so_j}.

The execution steps are as follows:

(1) The preset time length of the system scheduling cycle is $Sche_{time}$, and then the current system time is obtained through system calls as the start time of the scheduling task: T_{stime}.

(2) Calculate the P_{ltime} parameters of the startup time T_{stime} and P_{so_j} to obtain the scheduling priority. The formula is as follows: the larger the values of $T_{pri} = (T_{stime} - P_{ltime}) \div Sche_{time}$ and T_{pri}, the higher the priority and the higher the urgency of scheduling.

(3) Based on the network conditions and P_{size} parameters of P_{so_j}, assuming that the network state between Aso_i and P_{so_j} is in an ideal state, the minimum execution time of the scheduling task can be estimated, with $T_{exectime} = P_{size} \div (A_{net}|P_{net})$, A_{net}, and P_{net} taking the minimum value.

(4) The scheduling strategy uses a relatively simple weighted round robin scheduling algorithm, with T_{pri} as the weight, in order to schedule among the nodes in the edge computing model through polling. Based on the completion status of the scheduled task, set the value of $Task_{ij}$ as the logical value (Ture | False), where Ture indicates that the current task has been completed and False indicates that the scheduled task was not successful.

4. Result analysis.

4.1. Testing Environment. In order to reduce the impact on normal teaching order, a test is conducted before the summer vacation in a studio that is open to students all day. Based on the architecture diagram shown in Figure 4.1, the video capture framework is deployed using the design described in Section 3.5. ECU nodes are deployed, cameras are installed, the network is configured, and an IVSU is constructed. The relevant equipment and main parameters are shown in Table 4.1.

After completing the deployment of IVSU in the studio, the videos collected by IVSU devices are transmitted through wireless networks for data transmission; Edge computing node (ECN) is deployed in the equipment room on the floor; Then, it is transmitted to the Cloud Computing Data Center (CCDC) through the campus network, and the ECN device uses a DELL server. The main parameters are shown in Table 4.2.

During this testing process, the design and deployment of Cloud Computing Data Center (CCDC) was not carried out. A virtual machine (VM) was applied for in the existing cloud computing data center of the school, equipped with corresponding software environment, to simulate CCDC for data storage and scheduling functions. The parameters are shown in Table 4.3.

Table 4.1: Intelligent Video Monitoring Unit Equipment and Its Main Parameters

Serial Number	Equipment (software and hardware)	Parameters
1	Raspberry Pi Single Board Computer(Zere W)	CPU: 1 GHz, single core; Memory: 512 MB Network connection: 802.11 b/g/n; Storage interface: microSD camera interface: CSI; Power interface: micro USB
2	CSI camera (SONY IMX219)	Sensor: 8-megapixel CMOS Static image specifications: 3280 ∗ 2464 (maximum) Video recording specifications: ① 720p 60 fps; ② 1 080 p30 fps
3	storage device	MicroSD: 256G, Kingston KF-C38256-4K I/O performance: ① Read rate: 100 MB/s; ② Write rate: 80 MB/s
4	motionEyeOS	Release: 20190427; Motion: 4 2
5	Wireless router (TL WAR450L)	Transmission frequency band: 2.4 GHz; Transmission rate: 450 M LAN: Gigabit Ethernet port; Wireless gain: 5 dBi
6	exchange board	Huawei, S1700-24GR, 24 port

Table 4.2: Intelligent Video Monitoring Unit Equipment and Its Main Parameters

Option	Parameters/Model	Memo
Server	DELL PowerEdge R510	
CPU Xeon	X5650 * 2	2.60 GHz, single CPU with 6 physical cores, supporting hyper threading
Memory	32G	
Hard disk	1T * 4	Build a RAID-0 array to improve I/O performance
network	1 000 Mbps	
Disk alignment card	Dell PERC H700	512M cache
Operating system	CentOS 7 1810	Kernel version: 3 10 0 957 12 1; Docker version: 1.13.1; Python version: 2.7.5; PHP version: 5.4.16; Database version: MariaDB 5.5.60; Web server: Apache -2.4.6

4.2. Model testing.

4.2.1. Verification of motion detection function. Use IVSU to record a video. Extract the 1057th, 1058th, and 1059th frames for testing. After binarizing the above three frames of images. In the model designed by the author, a three frame difference algorithm is used for motion object detection. The unchanged ones are binarized and transformed into black backgrounds. The motion targets extracted by the algorithm have clear and complete contours, and the results meet the expected motion detection requirements.

4.2.2. Verification of storage requirements. According to the design in the third part, and also to simplify the management of storage files, the video stream is stored in time segments to record video information.

Table 4.3: Main parameters of cloud computing data centers

Option	Parameters/Model
CPU	Xeon E5 2609 v2(2 50 GHz) 8vCPU core
Memory	64 G
Hard disk	2 17TB
network	1 000 Mbps
operating system	CentOS 7 1810

Table 4.4: Video surveillance records

Week	Weekly			
	Week 1	Week 2	Week 3	Week 4
one	160	163	167	54
two	161	163	168	34
three	161	166	171	24
four	161	162	177	20
five	162	162	178	10
six	166	168	180	0
day	167	166	180	0

Take every 300 seconds as a time period, which is 5 minutes, as a video file. When a moving object is detected in the stored video frame, the status of the video recording file is marked and video information is stored. When there are no moving objects in a continuous 5-minute video frame, delete the video file without any motion state to save storage space. The standard opening hours of this studio are from 07:30 in the morning to 23:00 at night. Before deploying the IVSU, investigate the regularity of students entering the studio. Activities in the studio are mainly from 07:30 a.m. to 12:15 a.m. and from 14:10 p.m. to 23:00 p.m, there is less activity during the rest of the time. There is no activity time every day, with an average of about 10 hours and 25 minutes, accounting for approximately 43.30% of the entire day. After a 4-week test from June 10th to July 7th, the first to second weeks in Table 4.4 are the normal teaching weeks of the school; The third week is the pre exam review week; The fourth week is the exam week. The video surveillance records are shown in Table 4.4, where the data represents the number of files and each video file has a length of 5 minutes.

From the above test data, it can be seen that the number of video recording files is directly proportional to student activities. During the normal teaching week of the school, students enter and exit the studio and engage in regular activities within the studio. The amount of video storage remains relatively stable, with a slight increase towards the end of the semester. The third week corresponds to the pre exam review week, and the number of practical activities for students in the studio has increased. The fourth week is the exam week, and there is a significant decrease in activities within the studio. The holiday starts on Friday afternoon in the fourth week, and after the studio is closed, the number of video recording files is recorded as 0. Statistically analyze the data, compare the video surveillance with added motion detection function, and analyze the data from week 1 to week 3 to compare the video surveillance without added motion detection function. The number of video files stored is about 1176 records per week, saving about 41.56% of storage space. Therefore, if applied across the entire school, it can significantly reduce construction costs in terms of storage.

4.2.3. Verification of scheduling tasks. During the test, the intelligent video monitoring system has not been deployed in the whole school for the time being, and the scheduling verification focuses on the test from edge computing node (ECN) to edge computing unit (ECU) to detect the utilization of network bandwidth by the intelligent video monitoring system. Utilizing the lightweight and other features of the Docker container engine, by periodically collecting resource load information such as CPU, memory, disk I/O, and network bandwidth, information support is provided for containerized scheduling schemes. Assuming there are n ECU nodes in total, the ECU list is: $ecu = ecu_0, ecu_1, \cdots, ecu_{n-1}\{\}$, $weight(ecu_j)$ represents the weight of the jth

ECU node, that is T_{pri} calculated in formula 3.6, j also represents the object Pso_j scheduled last time, and $max(ecu)$ represents the maximum value among all nodes. $gcdnumber(ecu)$ represents the maximum common divisor of the weights of all nodes in the ECU list. The variable j is initialized to -1, represents the current weight, and initialized to 0. Increase the corresponding weights, determine the priority of each node, and avoid the data of a certain ECU node not being backed up for a long time. At the same time, through the corresponding weight, the load of network traffic and other conditions are fully considered to avoid the centralized scheduling of ECU nodes in the edge computing model, resulting in network congestion and reducing the overall network construction cost.

5. Conclusion. The scheme proposed by the author introduces the edge computing model to organically integrate the cloud computing data center and edge computing environment in the campus network. The open source and open hardware and software architecture are adopted to make full use of the computing resources of edge computing nodes to realize the motion detection function and effectively reduce the storage space requirements of the monitoring system; And use the Docker containerized platform to collect resource status information from each node, design resource scheduling strategies, and improve the utilization of network bandwidth. Intelligent video monitoring technology is a trend in the development of monitoring technology in the era of big data. Therefore, in subsequent work, the computing power of ECU will be combined to achieve target recognition and tracking functions in intelligent video monitoring systems; Optimize storage space, adopt distributed elastic storage mechanism, and fully utilize the storage capacity of ECN.

REFERENCES

[1] Rajavel, R., Ravichandran, S. K., Harimoorthy, K., Nagappan, P., & Gobichettipalayam, K. R. (2022). IoT-based smart healthcare video surveillance system using edge computing. Journal of ambient intelligence and humanized computing, 13(6), 3195-3207.

[2] Wan, S., Ding, S., & Chen, C. (2022). Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles. Pattern Recognition, 121, 108146.

[3] Xu, X., Wu, Q., Qi, L., Dou, W., Tsai, S. B., & Bhuiyan, M. Z. A. (2020). Trust-aware service offloading for video surveillance in edge computing enabled internet of vehicles. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1787-1796.

[4] Wang, F., Zhang, M., Wang, X., Ma, X., & Liu, J. (2020). Deep learning for edge computing applications: A state-of-the-art survey. IEEE Access, 8, 58322-58336.

[5] Yar, H., Imran, A. S., Khan, Z. A., Sajjad, M., & Kastrati, Z. (2021). Towards smart home automation using IoT-enabled edge-computing paradigm. Sensors, 21(14), 4932.

[6] Ke, R., Zhuang, Y., Pu, Z., & Wang, Y. (2020). A smart, efficient, and reliable parking surveillance system with edge artificial intelligence on IoT devices. IEEE Transactions on Intelligent Transportation Systems, 22(8), 4962-4974.

[7] Yu, J., Chen, M., Zhou, H., Wang, L., Luo, H., & Lan, J. B. (2022). Research on application of edge calculation in power grid state prediction. 2022 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST), 651-655.

[8] Song, W., Tang, Y., Tan, W., & Ren, S. (2023). Ishd: intelligent standing human detection of video surveillance for the smart examination environment. Computer Modeling in Engineering and Science (in English), 7(8), 6722-6747.

[9] Isaac Martín de Diego, Alberto Fernández-Isabel, Ignacio San Román, Conde, C., & Cabello, E. (2022). Novel context-aware methodology for risk assessment in intelligent video-surveillance systems. International Journal of Sensor Networks, 13(5), 118.

[10] Khosravi, M. R., & Samadi, S. (2022). Mobile multimedia computing in cyber-physical surveillance services through uav-borne video-sar: a taxonomy of intelligent data processing for iomt-enabled radar sensor networks. Tsinghua Science and Technology, 27(2), 288-302.

[11] Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457-7469.

[12] Lv, Z., Chen, D., Lou, R., & Wang, Q. (2021). Intelligent edge computing based on machine learning for smart city. Future Generation Computer Systems, 115, 90-99.

[13] Hua, H., Li, Y., Wang, T., Dong, N., Li, W., & Cao, J. (2023). Edge computing with artificial intelligence: A machine learning perspective. ACM Computing Surveys, 55(9), 1-35.

[14] Zhang, X., Cao, Z., & Dong, W. (2020). Overview of edge computing in the agricultural internet of things: Key technologies, applications, challenges. Ieee Access, 8, 141748-141761.

[15] Jiang, X., Yu, F. R., Song, T., & Leung, V. C. (2021). A survey on multi-access edge computing applied to video streaming: Some research issues and challenges. IEEE Communications Surveys & Tutorials, 23(2), 871-903.

[16] Guo, H., Liu, J., Ren, J., & Zhang, Y. (2020). Intelligent task offloading in vehicular edge computing networks. IEEE Wireless Communications, 27(4), 126-132.

[17] Qiu, T., Chi, J., Zhou, X., Ning, Z., Atiquzzaman, M., & Wu, D. O. (2020). Edge computing in industrial internet of things: Architecture, advances and challenges. IEEE Communications Surveys & Tutorials, 22(4), 2462-2488.

[18] Kong, X., Wang, K., Wang, S., Wang, X., Jiang, X., Guo, Y., ... & Ni, Q. (2021). Real-time mask identification for COVID-19: An edge-computing-based deep learning framework. IEEE Internet of Things Journal, 8(21), 15929-15938.

[19] Zhou, X., Xu, X., Liang, W., Zeng, Z., & Yan, Z. (2021). Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart IoT. IEEE Internet of Things Journal, 8(16), 12588-12596.

[20] Hartmann, M., Hashmi, U. S., & Imran, A. (2022). Edge computing in smart health care systems: Review, challenges, and research directions. Transactions on Emerging Telecommunications Technologies, 33(3), 3710.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 10, 2024
Accepted: Jul 19, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1242–1256, DOI 10.12694/scpe.v26i3.4130

THE APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGY IN HUMAN CENTERED MANUFACTURING IN INDUSTRY 5.0

JIAWEI ZHANG*

Abstract. In order to clarify the cognitive process of human beings and the influencing factors of human errors in the process of manufacturing capability evaluation, quantitatively analyze the reliability of human beings in the manufacturing process, and more accurately evaluate the manufacturing capability of production lines, the author proposes the application of artificial intelligence technology in human centered manufacturing in Industry 5.0. In response to the dynamic nature of data in the operation process of manufacturing production units and the varying importance of indicators to evaluation objects at different times, the author proposes an objective weighting method that combines indicator sensitivity with entropy weight method to solve the problem of existing weighting methods only considering the fluctuation of indicator data and ignoring the importance of evaluation indicators to all evaluated objects. The combination of subjective weights established by the Analytic Hierarchy Process (AHP) is used to obtain the final combination weight of indicators. At the same time, evaluate and analyze the factors that contribute to human error behavior to obtain the human reliability of the unit, and introduce it into the comprehensive evaluation of unit manufacturing capability. Based on the time series data in the evaluation, a time dimension factor combined with grey correlation analysis is introduced to conduct a dynamic comprehensive evaluation of unit manufacturing capacity in time series, and the production unit manufacturing capacity index is obtained. The example results show that 10 indicator data from the past 10 time periods were selected for evaluation, and the closer the time period, the more important the data is. The time factor for each time period is (0.0048, 0.0068, 0.0126, 0.0266, 0.0582, 0.0704, 0.1232, 0.1685, 0.2212, 0.3070). The unit capability value obtained through dynamic horizontal and vertical comprehensive evaluation is most consistent with the capability value obtained by the author's method. Under the four methods, although there are differences in the capacity values of each unit, the fluctuation is within a reasonable range, indicating that the author's evaluation method is reasonable and feasible. The feasibility and effectiveness of the evaluation method have been validated.

Key words: Human reliability, Manufacturing capability, Dynamic evaluation, Production line, Industry 5.0

1. Introduction.

Human centric/centered manufacturing, also known as people-oriented manufacturing, abbreviated as human-centered manufacturing, mainly involves two subjects - humans and machines, as well as their relationship - human-machine relationship . With the emergence and development of new generation information and communication technology (ICT)/artificial intelligence (AI) technologies such as the Internet of Things, cloud computing, Cyber physical systems (CPS), big data, and deep learning, the arrival of Industry 5.0, which mainly relies on intelligent manufacturing, has been promoted. New types of operators - operator 5.0 and holographic perception intelligent connected autonomous intelligent machines have emerged. The human-machine relationship (especially human-machine interaction) has evolved from the initial physical direct interaction between a single person and a single machine to the system collaboration of virtual and real fusion between humans and objects. Human in the loop (HiL) is no longer limited to the physical loop, and concepts such as Human on the loop (HoL) and Human out of the loop (HofL) have emerged [1,2]. In fact, HiL, HoL, and HofL correspond to the physical space, information space, and social (community) space of human intelligent manufacturing, ultimately forming a trinity of autonomous social cyber physical production system (SCPPS); Especially the development of intelligent manufacturing in Human CPS (H-CPS) promotes the development of people-oriented intelligent manufacturing [3].

Compared with the world's advanced level, the manufacturing industry has problems such as being large but not strong. The gap in independent innovation ability, resource utilization efficiency, industrial structure level, informatization level, quality and efficiency is particularly obvious. The task of industrial intelligence transformation and upgrading and leapfrog development is urgent and arduous [4]. The continuous development

*School of Art and Design, Luoyang Vocational College of Science and Technology, Luoyang, Henan, 471822, China. (Corresponding author, JiaweiZhang7@163.com)

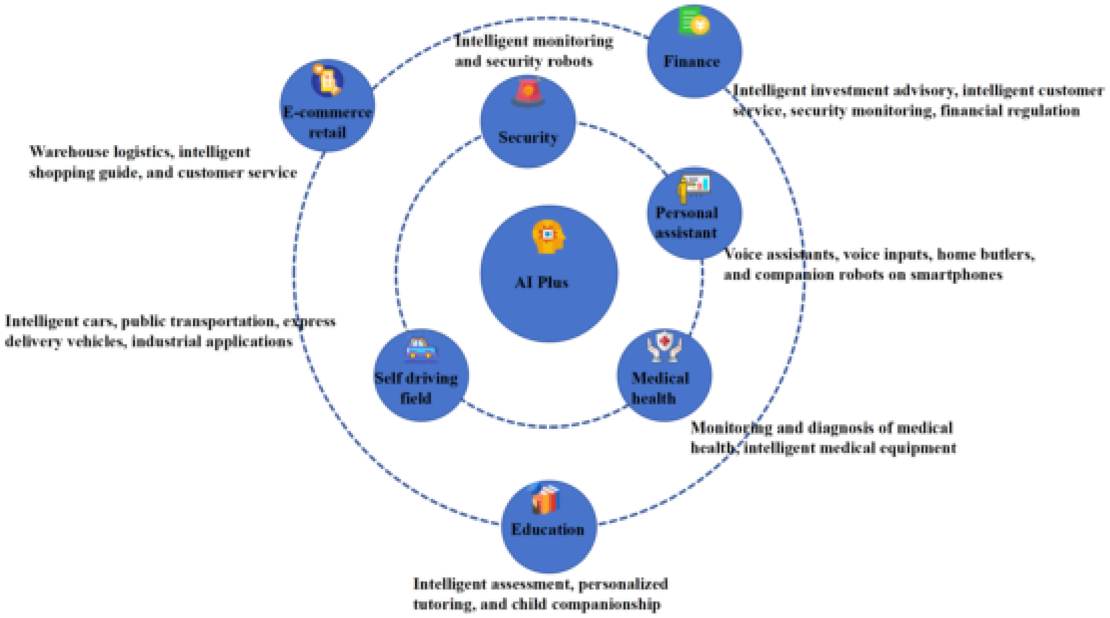

Intelligent monitoring
and security robots
Finance
E-commerce
retail
Intelligent investment advisory, intelligent customer
service, security monitoring, financial regulation
Security
Warehouse logistics, intelligent
shopping guide, and customer service
Personal
assistant
AI Plus
Voice assistants, voice inputs, home butlers,
and companion robots on smartphones
Intelligent cars, public transportation, express
delivery vehicles, industrial applications
Self driving
field
Medical
health
Monitoring and diagnosis of medical
health, intelligent medical equipment
Education
Intelligent assessment, personalized
tutoring, and child companionship

Fig. 1.1: Artificial Intelligence Technology

of new generation Internet, artificial intelligence, digital twins and other technologies has continuously injected strong power into the development of intelligent manufacturing. The excessive pursuit of informatization and digitization in production models can no longer meet the needs of complex operations such as flexible production workshops and personalized customization of users. The difficulties in intelligent manufacturing are becoming more prominent, so the production trend urgently needs to be changed, and human beings as a key factor cannot be ignored anymore [5]. The concept of Industry 5.0 has gradually attracted people's attention. As a continuation and supplement to Industry 5.0, Industry 5.0 not only focuses on optimizing industrial structure and improving automation levels, but also places people at the center of the manufacturing industry, allowing technology to actively serve and adapt to people, and paying more attention to human values and feelings. Human centered intelligent manufacturing should consider the safety and happiness of workers, dispel their concerns and concerns about the "machine replacement" brought about by the industrial revolution, and allow labor to return to the factory [6] (Figure 1.1).

2. Literature Review. In the blueprint of intelligent manufacturing, human-machine collaboration has become the mainstream mode of production and service. Due to the deep collaboration between humans and machines, the tasks and requirements of humans in intelligent manufacturing systems have undergone significant changes. Although humans no longer bear repetitive tasks, they remain the central link of the decision-making loop system and always occupy a dominant position. The deep connotation of human-machine collaboration is the integration of human-machine intelligence, which represents the need for humans and machines to jointly complete designated tasks. In the process of completing dynamic job tasks, the manufacturing system needs to keep pace with the staff, face dynamic job requirements, adapt resources and cooperate autonomously to achieve coordinated production. The intelligent manufacturing development theory of human cyber physical system (HCPS) proposed by Rannertshauser, P., clarifies a technological system dominated by physical systems (machines, robots, processing processes), information systems, and human decision-making [7]. By transferring part of human perception, analysis, and control functions through information systems, it can replace most of human physical and mental labor.

Compared with the automotive industry, manufacturing tasks and processes in fields such as aerospace, shipbuilding, and construction are too complex and require high assembly accuracy. Currently, manual operations are still relied upon, so it is necessary to conduct research on human-machine collaboration. Assisting humans with collaborative robots to complete complex tasks, maintaining optimal levels of mental and physical strength, and balancing the demand for cognitive resources in the brain with the supply of cognitive resources to the task, avoiding the negative impact of overload and underload on operators, reducing human burden, improving task execution efficiency and production safety. Froschauer, R. studied the relevant algorithms for controlling cooperative robots using electromyographic signals [8]. Romero, D. explored the effectiveness of guiding gestures in human-machine collaboration scenarios in the industrial field [9]. The "Intelligent Unit Production Line for Human Machine Collaboration" launched by Romero D integrates advanced technologies such as artificial intelligence, Internet of Things, and big data analysis. It aims to improve production efficiency, flexibility, and quality for multi variety and small batch production modes. Through mutual perception between humans and machines, ultra flexible production can be achieved on the same site through complementarity and assistance [10]. The "Intelligent Flexible Production Line Using Robots to Produce Robots" launched by Zhang R provides a stable and efficient reference sample for the human-machine cooperation application of collaborative robots in the industrial production field through modular design and collaborative production methods [11].

The author studies a dynamic evaluation method for unit level manufacturing capability based on information sensitivity at the unit level. Firstly, a production unit manufacturing capability evaluation model is established, and then human reliability issues are considered in the unit capability evaluation. Static grey correlation analysis is extended to dynamic decision-making, and indicator sensitivity weights are used to modify and assign weights, time dimension factors are introduced to obtain the unit's various capability values and comprehensive manufacturing capability values at each time step, based on time series data, the total manufacturing capability of each unit is obtained.

3. Research Methods.

3.1. Dynamic evaluation model for manufacturing capacity at the production unit level. During the production process of manufacturing system production units, each processing state will change with time, and the evaluation index values are also constantly changing. Therefore, the capacity value of the unit is dynamically changing at different times [12]. The static evaluation method mainly involves a two-dimensional evaluation of the decision object at a single moment, which only includes the decision space and the target space, and cannot reflect the characteristics over a period of time. Therefore, in addition to evaluating the decision space and target space dimensions, the production unit level manufacturing capacity also needs to be extended to consider time and space, that is, to dynamically evaluate the production unit level manufacturing capacity from the three dimensions of time, indicators, and goals [13].

In the process of evaluating unit capabilities, the focus is on the establishment of evaluation indicators and the generation of manufacturing capability evaluation results. In order to dynamically evaluate unit level capabilities, it is necessary to consider the indicator data of the entire time series. Let the decision solution set $U = \{u_1, u_2, \cdots, u_n\}$ consist of n manufacturing unit objects to be evaluated, the indicator set $P = \{p_1, p_2, \cdots, p_m\}$ consists of m evaluation indicators. The manufacturing indicator data from nearly N time points in the unit manufacturing process is used as the evaluation basis data. If the jth attribute value of the evaluation unit u_i at time point $t_k(k = 1, 2, \cdots, N)$ (or stage) is $p_{ij}(t_k)$, then the unit evaluation indicator data chronological list can be formed as shown in Table 3.1.

In the process of dynamically evaluating the manufacturing capacity of production units, it is necessary to first conduct a two-dimensional static evaluation at a fixed time, and then comprehensively evaluate the static evaluation results in the time dimension. For the evaluation of dynamic 3D space, commonly used 3D evaluation operators include Time Order Weight Average operator (TOWA) and Time Order Weighted Geometric Average operator (TOWGA) [14]. The time-series weighted average operator first evaluates at each time step, and then evaluates the time dimension; The temporal geometric mean operator evaluates the time dimension and then reduces it to evaluate the indicator dimension. The author uses a time-series weighted average operator to evaluate the manufacturing capacity of production units in three-dimensional space.

Based on the index time sequence list established in Table 3.1, at time $y_i(t_k) = \sum_{j=1}^{m} a_j p_{ij}(t_k), k = 1, 2, \cdots, n$, the comprehensive evaluation function of production unit iu based on the weighted model can be

Table 3.1: List of Index Time Series

unit	t_1				t_2				$\cdots$	t_N			
	$p_1, p_2, \cdots, p_m$				$p_1, p_2, \cdots, p_m$				$\cdots$	$p_1, p_2, \cdots, p_m$			
u_1	$p_{11}(t_1)$	$p_{12}(t_1)$	$\cdots$	$p_{1m}(t_1)$	$p_{11}(t_2)$	$p_{12}(t_2)$	$\cdots$	$p_{1m}(t_2)$		$p_{11}(t_N)$	$p_{12}(t_N)$	$\cdots$	$p_{1m}(t_N)$
u_2	$p_{20}(t_1)$	$p_{21}(t_1)$	$\cdots$	$p_{2m}(t_1)$	$p_{20}(t_2)$	$p_{21}(t_2)$	$\cdots$	$p_{2m}(t_2)$	$\cdots$	$p_{20}(t_N)$	$p_{21}(t_N)$	$\cdots$	$p_{2m}(t_N)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
u_n	$p_{n1}(t_1)$	$p_{n2}(t_1)$	$\cdots$	$p_{nm}(t_1)$	$p_{n1}(t_2)$	$p_{n2}(t_2)$	$\cdots$	$p_{nm}(t_2)$		$p_{n1}(t_N)$	$p_{n2}(t_N)$	$\cdots$	$p_{nm}(t_N)$

described as follows 3.1:

$$y_i(t_k) = \sum_{j=1}^{m} \alpha_j p_{ij}(t_k), k = 1, 2, \cdots, N; i = 1, 2, \cdots, n \tag{3.1}$$

In the formula, $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$ is the weight corresponding to each evaluation indicator, and it satisfies the following equation 3.2:

$$\sum_{j=1}^{m} \alpha_j = 1, 0 \leqslant \alpha_j \leqslant 1 \tag{3.2}$$

For the calculation of indicator weights, it is usually necessary to pay attention to both people's subjective experience information and the information brought by objective data itself. Therefore, a combination of subjective weight and objective weight is used to assign weights to the indicators. If the subjective weight value is w' and the objective weight value is w", the combination weight of the indicators is as follows 3.3:

$$\alpha_j = \frac{w'_j * w_j"}{\sum_{j=1}^{m} w'_j * w_j"} (j = 1, 2, \cdots, n) \tag{3.3}$$

The author uses the sensitivity weight of indicators to adjust the objective weight, as the commonly used objective weighting method often only focuses on the fluctuation of the indicator's own data and does not pay attention to the impact of indicator changes on the overall indicator set. In order to more accurately reflect the weight of indicators, the author uses the sensitivity weight of indicators to adjust the objective weight. If the sensitivity weight of indicators is , the modified objective weight is as follows 3.4:

$$w_j" = \frac{w_j * w_r}{\sum_{j=1}^{m} w_j * w_r} \tag{3.4}$$

Due to the fact that the value of unit manufacturing capacity varies at different times in the time sequence, and the proportion of unit manufacturing capacity to the total manufacturing capacity of the production unit at different times in the time sequence is not the same [15]. Due to the influence of interference factors at certain times, the unit evaluation index data may experience abnormal mutations. In this case, the unit manufacturing capacity evaluation results obtained do not match the actual situation. Therefore, in order to avoid this situation, after conducting a two-dimensional static evaluation, different time factors can be assigned to the evaluation values at different times for comprehensive consideration, by using a time factor, the ability values at each time step are integrated into the total ability value at the time step. If the time factor at each moment is set to $v = \{v_1, v_2, \cdots, v_n\}$, then v should satisfy the following equation 3.5:

$$\sum_{j=1}^{N} v_j = 1, 0 \leqslant v_j \leqslant 1 \tag{3.5}$$

Therefore, the manufacturing capacity value of production unit v_i in the time sequence can be expressed as equation 3.6:

$$y_i = \sum_{k=1}^{N} y_i(t_k) * v_k \tag{3.6}$$

Based on the above research content, it can be concluded that for the dynamic evaluation of manufacturing capacity of production units, a two-dimensional evaluation should be conducted first in the time series table, and then a comprehensive manufacturing capacity evaluation of the unit should be conducted in the time dimension. In this process, the production data of the unit in the manufacturing process is updated in real-time in the time series table, which is used as a new data source for evaluating the manufacturing capacity of the unit. Therefore, the manufacturing capacity of the unit can be evaluated in real-time and dynamically. In order to establish a dynamic evaluation model for the manufacturing capacity of production units, providing support for subsequent dynamic evaluation methods.

The principle of the dynamic evaluation model for manufacturing capacity of production units is to first study and analyze the state change information during the manufacturing process of production units, including the status information of processing personnel during the manufacturing process of units. Based on the historical manufacturing task data of units, the evaluation data is standardized to form a unit capacity evaluation matrix [16]. The unit manufacturing data changes dynamically over time, and the importance of each indicator data to the overall evaluation indicator set may vary at different times, which may result in information redundancy. Therefore, the sensitivity of the indicator information is used to reflect the degree of influence of each indicator on the original indicator set information. Based on this, the indicator set is reduced in dimensionality, reducing redundant information, making the evaluation indicators more accurate, and further obtaining the sensitivity weight of the indicators. The sensitivity weight is used to correct the indicator weight, so that when assigning weights to indicators, not only the fluctuation of the indicator's own data is considered, but also the degree of influence of each indicator on the original indicator set information, at the same time, pay attention to the impact of indicator changes on the overall indicator set, obtain indicator weights, and combine grey correlation analysis to obtain the capability values of each unit's time series. Further introduce time factors, and finally obtain the total manufacturing capability values of each unit.

3.2. Optimization and weighting of evaluation indicators based on information sensitivity.

3.2.1. Optimization of evaluation indicators based on information sensitivity. In the evaluation of manufacturing capacity of production units, the data of evaluation indicators is constantly changing. During the evaluation process, a large amount of historical data needs to be combined for evaluation. The importance of each indicator data to the evaluation object may change at different times, and different indicator data may reflect the same information. That is, there may be information overlap between indicators, resulting in information redundancy. Some data information may be repeatedly emphasized during the evaluation, which may distort the evaluation results. Therefore, in the evaluation process, in order to avoid wasting calculation time on unnecessary data and make the evaluation results more accurate, it is necessary to reduce the dimensionality of the indicators.

Currently, principal component analysis, factor analysis, and other dimensionality reduction methods are still widely used in various fields such as comprehensive evaluation and pattern recognition. However, these methods still have some problems, such as difficulty in determining the economic meaning of principal components and non unique factor loading matrices. Moreover, most of these dimensionality reduction methods have not taken into account the degree of influence of indicators on the overall indicator set information, and may lose some information that has a significant impact on the overall indicator set during the dimensionality reduction process [17]. Therefore, based on the information sensitivity of the indicators, the author optimizes the dimensionality of the evaluation indicators to ensure that the retained indicator information has a significant impact on the original indicator set information and the degree of information overlap between the evaluation indicator sets is relatively low. Information sensitivity reflects the degree to which a certain indicator affects the information of the original indicator set. The greater the information sensitivity, the more important the indicator is in the original indicator system, and correspondingly, the more significant its impact on the evaluation results; On the contrary, it indicates that changes in indicators have a smaller impact on the evaluation results.

The dimensionality reduction of indicator data based on the information sensitivity of indicators is developed on the basis of principal component analysis dimensionality reduction method. The standardized data matrix of indicators is set as $X = (x_{ij})_{n\times m}$, among them, n represents the amount of indicator data, m represents the number of indicators, andamong them, n represents the amount of indicator data, m represents the number of indicators, and x_{ij} is the i-th data of the j-th indicator. The steps to reduce the dimensionality of the indicator using information sensitivity are as follows.

(1). Solve the principal component Z_i as follows 3.7:

$$Z_i = u_{i1}X_1 + u_{i2}X_2 + \cdots + u_{ij}X_j + \cdots + u_{im}X_m \tag{3.7}$$

In the formula, Z_i represents the i-th principal component, $X_j = (x_{1j}, x_{2j}, \cdots, x_{ij}, \cdots, x_{nj})$ is the value of the j-th indicator after Z-normalization of the indicator data, and u_{ij} is the j-th component of the orthogonal unitary eigenvector $u_i^T = (u_{i1}, u_{i2}, \cdots, u_{im})$ of the indicator correlation coefficient matrix X^TX.

(2). Calculate the variance contribution rate ω_i of principal component Z_i, as shown in equations 3.8 and 3.9:

$$|X^TX - \lambda_i E_m| = 0 \tag{3.8}$$

$$\omega_i = \lambda_i / \sum_{i=1}^{m} \lambda_i \tag{3.9}$$

Obtain the eigenvalues λ_i of the correlation coefficient matrix X^TX through equation 3.9, and the variance contribution rate ω_i reflects the proportion of the information content of the i-th principal component Z_i to the information content of all original indicators.

(3). Calculate the cumulative variance contribution rate Ω_k as follows 3.10:

$$\Omega_k = \sum_{i=1}^{k} \omega_i \tag{3.10}$$

In equation 3.10, k represents the number of retained principal components. Usually, in principal component analysis, several principal components with a cumulative variance contribution rate of 70%∼90% and higher information content are retained. In order to approximate the original indicator set information, the author selects the relatively higher proportion of 90%, therefore, if the cumulative variance contribution rate of the first k principal components is $\geqslant \Omega_k 90\%$, the top k principal components with the highest variance contribution rate are retained.

(4). Solve the information sensitivity β_j of the jth indicator as follows 3.11:

$$\beta_j = \sum_{i=1}^{k} \omega_i |\partial Z_i / \partial X_j| \tag{3.11}$$

In the formula, $|\partial Z_i / \partial X_j|$ represents the sensitivity of the i-th principal component to changes in the size of the j-th indicator, that is, the magnitude of the change in information content of the i-th principal component caused by a small change in the j-th indicator, while the size of other indicators remains unchanged.

(5). Calculate the cumulative information content Γ_s, which reflects the cumulative content of the reduced indicators in the original indicator set. Assuming that the sensitivity of indicator information is arranged in descending order, the result is $\beta_1^* > \beta_2^* > \cdots > \beta_m^* >$, then s Γ solution is shown in formula 3.12:

$$\Gamma_s = \sum_{j=1}^{s} \beta_j^* / \sum_{j=1}^{m} \beta_j \tag{3.12}$$

Table 3.2: Description of Priority Relationship Matrix

A	B_1	B_2	$\cdots$	B_m
B_1	B_{11}	B_{12}	$\cdots$	B_{1m}
B_2	B_{20}	B_{21}	$\cdots$	B_{2m}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
B_m	B_{m1}	B_{m2}	$\cdots$	B_{mm}

Generally, selecting to retain indicator information that can reach a cumulative information content of 70% to 90% for dimensionality reduction of the indicator set.

Reducing the dimensionality of indicators by determining their information sensitivity can compensate for the problem in principal component analysis where the load coefficient cannot reflect the importance of the indicators to the original indicator set, and reducing the dimensionality of information solely based on the load coefficient may be unreasonable [18]. By performing dimensionality reduction on indicator data, the pressure of evaluation calculations can be reduced without losing the original indicator set information, while further improving the final evaluation results.

3.2.2. Weighting of Evaluation Indicator Combination for Sensitivity Correction.

(1) Determination of subjective weight. In the process of weighting indicators, it is necessary to consider the importance of each indicator to the evaluation objectives based on the actual situation. Although subjective weighting carries a significant personal subjective color and the given weights may not conform to the changes in objective data, the evaluation process cannot ignore the decision-maker's view on the importance of the indicators. The subjective weighting method can effectively avoid the phenomenon of "important indicators having smaller weights and unimportant indicators having larger weights" that may occur in absolute objective weighting based on actual subjective conditions. Therefore, subjective weighting is crucial in the process of weighting indicators. The Analytic Hierarchy Process (AHP) is widely used in decision weighting, which is a weighting method that quantifies qualitative analysis problems by mathematizing people's thinking processes about complex systems [19]. The author subjectively assigns weights to unit capability evaluation indicators using the Analytic Hierarchy Process.

The Analytic Hierarchy Process (AHP) can transform a complex decision-making problem into a ranking problem of the evaluated object relative to the evaluation target. It first refines a complex decision-making problem into some constituent factors, and then further subdivides these factors until they cannot or do not need further subdivision. By doing so, a hierarchical structure can be established between various factors based on their subordinate relationships.

It can be specifically divided into the following steps:

Step 1: Establish a hierarchical structure. Based on the evaluation of unit manufacturing capability, the various factors that affect manufacturing capability are subdivided, and a hierarchical structure is formed based on the subordinate relationship between the factors [20].

In a hierarchical structure, there are generally three layers: target layer, criterion layer, and indicator layer. The targets of the two layers of indicators are often only related to some indicators in the lower layer, and the weights between indicators that do not have a connection between the upper and lower layers are 0.

Step 2: Establish a priority relationship matrix. Based on the established indicator hierarchy, establish the priority relationship matrix, also known as the judgment matrix, for the evaluation system. Compare the corresponding indicators in the upper and lower layers in sequence. The priority relationship matrix formed by the target layer and the criterion layer can be described in Table 3.2.

In the Table, B_{ij} represents the importance ratio scale of the indicators in criterion layer B relative to target layer A. Its value is generally determined using the 1-9 scale method based on people's intuitive judgment, as shown in Table 3.3, which shows the value pattern of B_{ij}. When B_{ij} takes numbers such as 2, 4, 6, and 8, it indicates that its importance is the middle value of adjacent levels.

Step 3: Consistency verification. After obtaining the judgment matrix, in order to make the final evaluation result reasonable, it is necessary to perform consistency judgment on it, and the judgment formula is shown in

Table 3.3: Proportional Scale of Relative Importance

Importance Level	B_{ij} value	Importance Level	B_{ij} value
i,j both elements are equally important	2		
Element is slightly more important than element j	4	The i element is slightly than the j element	1/4
The i element is significantly more important than the j element	6	The i element is significantly less important than the j element	1/6
The element i is much more important than the element j	8	The element i is much less important than the element j	1/8
The element i is extremely important than the element j	10	The element i is less important than the element j	1/10

Table 3.4: Average Random Consistency Index

Matrix order	1	2	3	4	5	6	7	8
RI	0	0	0.53	0.88	1.13	1.25	1.35	1.43
Matrix order	9	10	11	12	13	14	15	
RI	1.45	1.48	1.53	1.55	1.57	1.59	1.60	

equation 3.13.

$$CI = \frac{\lambda_{max} - m}{m - 1} \tag{3.13}$$

λ_{max} represents the maximum eigenvalue of the judgment matrix. If $\lambda_{max} = m$, that is CI=0, it indicates that the judgment matrix is completely consistent; On the contrary, if $CI \neq 0$, it indicates poor consistency. But this calculation process will become increasingly complex as the order of the judgment matrix increases. In order to make the calculation process relatively easy, the concept of random consistency ratio is introduced, it is the ratio of CI to RI, denoted as CR. The RI values of the 15th to 1st order judgment matrix are shown in Table 3.4. The following equation 3.14:

$$CR = \frac{CI}{RI} \tag{3.14}$$

If CR>1.0, then certain modifications should be made to the judgment matrix, and the above steps should be repeated until CR<0.10 is met, so there is no need to modify the judgment matrix.

Step 4: Sort by level. If CR<0.10, it is necessary to calculate the weight values of the relative importance order between the indicators in this layer and those in the previous layer in the judgment matrix.

Step 5: Overall hierarchical sorting. Repeat the operations from step 1 to step 4, from the top layer to the bottom layer, and calculate the eigenvalues and eigenvectors of each judgment matrix for each layer in turn, following the hierarchical structure. Then calculate the overall ranking of the hierarchy and calculate the relative weights of all indicators in the lowest layer relative to the target layer.

When the total ranking of all indicators $B_1, B_2, \cdots, B_m$ in the criterion layer is completed, the weights obtained are $b_1, b_2, \cdots, b_m$, if the single ranking result of a certain indicator c_j in layer C with respect to a certain indicator B_i in layer B is $c_1^i, c_2^i, \cdots, c_j^i, \cdots, c_n^i$, then the total ranking in layer C can be expressed as follows 3.15:

$$c_j = \sum_{i=1}^{m} b_i c_j^i (j = 1, 2, \cdots, n) \tag{3.15}$$

The overall hierarchical sorting still requires consistency testing. By following these steps and following the calculation steps above, the subjective weight value of the evaluation index can be obtained, which is denoted as W'.

(2) Sensitivity correction objective weight. Subjective weights often reflect the subjective preferences of decision-makers and cannot reflect the changes in objective data of indicators. If a certain indicator is very important but its data changes in each period are minimal, the impact on the evaluation object is relatively small; If a certain indicator has a low level of importance but has significant differences in data changes across different periods, it may have a significant impact on the evaluation results, therefore it needs to be given a higher weight. As an objective weighting method, the entropy weighting method determines the weight of each indicator based on the amount of information provided by its entropy value. The core idea is to reflect the importance of a certain indicator based on the degree of difference between its observed values. If the data difference of a certain indicator of each evaluated object is not too large, it indicates that the indicator has little effect on the evaluation system [21]. So it is inversely correlated with entropy value, while the importance of an indicator is positively correlated with entropy weight, that is, the larger the entropy weight, the more important the indicator is; On the contrary, the smaller the entropy weight, the less important the indicator is. Compared with other evaluation methods, entropy weighting method can avoid the interference of human factors in the weighting process of indicators, making the evaluation results more realistic.

Using the entropy weight method to assign weights to indicators, the specific steps are as follows:

Let $X = (x_{ij})_{n\times m}$ be the data matrix of n evaluation sequences and m preprocessed evaluation indicators. Let H_j be the entropy value of the jth evaluation indicator, and then the entropy value H_j is as follows 3.16, 3.17:

$$H_j = -\frac{1}{ln(n)}(\sum_{i=1}^{n} f_{ij} ln f_{ij})(i = 1, 2, \cdots, n, j = 1, 2, \cdots, m) \tag{3.16}$$

$$f_{ij} = \frac{u_{ij}}{\sum_{i=1}^{n} u_{ij}} \tag{3.17}$$

According to the entropy value obtained above, the weight value of the indicator can be obtained as follows 3.18:

$$w_j = \frac{1 - H_j}{\sum_{j=1}^{m}(1 - H_j)}(j = 1, 2, \cdots, m) \tag{3.18}$$

The entropy weight method focuses on the magnitude of the fluctuation of the indicator's own data, with large data fluctuations indicating a greater impact on the evaluation results and small data fluctuations indicating a smaller impact on the evaluation results. However, it often overlooks the degree of influence of a certain indicator on the entire evaluation indicator set. The sensitivity of indicator information mentioned in the previous section reflects the degree of impact of indicator changes on the overall evaluation indicator set. Therefore, the weighting of indicators is based on their information sensitivity, and the calculation formula is shown in equation 3.19.

$$w_r = \beta_r / \sum_{j=1}^{m} \beta_j \tag{3.19}$$

After obtaining the sensitivity weights of the indicators, the weights obtained by the entropy weight method are modified to obtain more accurate objective weights of the indicators, significantly improving the accuracy of the evaluation results. Using the multiplication integration normalization method to combine the two, the calculation formula is as follows 3.20:

$$w_j" = \frac{w_j * w_r}{\sum_{j=1}^{m} w_j * w_r} \tag{3.20}$$

(3) Combination weighting. In the evaluation of manufacturing capacity in production units, the weight of indicators mainly includes two aspects: On the one hand, it is subjective weighting, which assigns weights to indicators by quantifying the decision-maker's subjective preference for the indicators; On the other hand, the weighting of indicators is based on their objective data, reflecting the usefulness of the amount of information that the objective data can provide during the evaluation process. Therefore, when assigning weights to indicators, one should not only consider one aspect of the weight, but should combine subjective and objective factors for combined weighting. The author uses a aggregation method with multiplication characteristics to aggregate subjective and objective weights, and obtains the final weight of the aggregated indicators as $w = (w_1, w_2, \cdots, w_m)$. Among them, equation 3.21 is as follows:

$$w_j = w'_j * w_j" / \sum_{j=1}^{m} w'_j * w_j" (j = 1, 2, \cdots, n) \tag{3.21}$$

4. Result analysis.

4.1. Example verification. A certain manufacturing workshop has six production and processing units u_1, u_2, u_3, u_4, u_5, and u_6 responsible for the production and processing of workshop tasks. The production unit evaluation index system established earlier evaluates the manufacturing capacity of production units from the aspects of processing quality Q, processing flexibility F, manufacturing time T, processing cost C, environmental protection E, and human reliability P, selecting production and manufacturing data from nearly ten time periods of each unit to form the original unit evaluation index time series data Table, according to the author's proposed unit manufacturing capability evaluation method, the specific evaluation steps for its manufacturing capability are as follows.

Step 1. In the unit capability evaluation, the evaluation indicators are both benefit indicators and cost indicators. Formulas 3.1 and 3.2 are used to preprocess the original indicator data to obtain a standardized temporal decision matrix, which provides data support for subsequent evaluations [22].

Step 2. For the preprocessed indicator data, the impact of each indicator on the overall evaluation indicator set varies at different times, taking the most recent 10t as an example, calculate the information sensitivity B_j of each indicator according to formulas 3.7 to 3.11, and obtain the cumulative information content Γ_s of each indicator from formula 3.12. Then, perform dimensionality reduction on the indicators. The calculation results are shown in Table 4.1.

Generally, retaining indicator information with a cumulative information content of 65% to 95% can be achieved. The author chooses to retain information with a cumulative information content of 95%. According to Table 4.2, after sorting by sensitivity, the cumulative information content at indicator E3 reaches 88%. Therefore, at 10t, retaining indicator E3 and its previous indicator information is used to reduce the dimensionality of the indicator set.

Step 3. For the dimensionality reduction processed indicators, calculate the sensitivity weights of each indicator according to formula 3.19 as shown in Table 4.2.

Step 4. Based on the Analytic Hierarchy Process, subjectively assign weights to each indicator, and use formulas 3.16 to 3.20 to modify the entropy weights of each indicator using sensitivity weights to obtain the objective weights of the indicators. Finally, according to formula 3.21, obtain the combined weights of each indicator in indicator layer C relative to criterion layer B and relative to target layer A, as shown in Table 4.3.

Step 5. During the evaluation process, 10 indicator data from the past 10 time periods were selected for evaluation. The closer the time period, the more important the data is often, taking λ=0.4 here, we can obtain the time factor for each time step as v=(0.0048,0.00680.0126,0.0266,0.0582,0.0704,0.1232,0.1685,0.2212,0.3070)

Step 6. Determine the target sequence of each indicator, in order to obtain the capability values of each unit in terms of quality, time, and comprehensive manufacturing capability at each time step [23].

After obtaining the capability values at each moment, combined with the time factors obtained earlier, the total capability values in terms of unit processing quality, manufacturing time, etc., as well as the total manufacturing capability values, are shown in Table 4.4.

Table 4.1: Sensitivity of Indicator Information

Index	Sensitivity B_j	Sort by B_j	Accumulated information content Γ_s
Q1	0.297	Q1(0.299)	6.32%
Q2	0.16	F3(0.286)	12.32%
Q3	0.218	T2(0.274)	19%
Q4	0.224	T3(0.265)	23.61%
C1	0.255	C1(0.255)	28%
C2	0.156	T1(0.243)	35%
C3	0.172	Q4(0.225)	38.75%
F1	0.186	Q3(0.218)	43.32%
F2	0.208	P6(0.214)	47.81%
F3	0.286	F2(0.208)	52.11%
T1	0.242	P3(0.202)	56.41%
T2	0.275	P4(0.197)	60.51%
T3	0.267	E2(0.193)	64.51%
El	0.182	P2(0.188)	68.41%
E2	0.192	F1(0.186)	72.31%
E3	0.168	E1(0.183)	76.11%
P1	0.164	PS(0.175)	79.81%
P2	0.188	C3(0.175)	83.41%
P3	0.202	E3(0.167)	88%
P4	0.197	P1(0.164)	90.41%
P5	0.176	C2(0.157)	93.71%
P6	0.214	P7(0.153)	96.81%
P7	0.153	Q2(0.152)	100%

Table 4.2: Indicator Sensitivity Weights

Index	Q_1	Q_2	Q_3	C_1	C_2	F_1	F_2
Sensitivity weight	0. 404	0. 295	0.302	0. 598	0. 403	0. 274	0. 307
index	F_3	T_1	T_2	T_3	E_1	E_3	E_2
Sensitivity weight	0. 423	0. 308	0. 352	0. 343	0.337	0. 423	0.354
index	E_3	P_2	P_3	P_4	P_5	P_6	
Sensitivity weight	0. 313	0. 193	0. 208	0. 203	181	0. 218	

4.2. Result Analysis. The author first optimized the dimensionality of indicators based on their information sensitivity. The degree of influence of indicator data on the overall evaluation object may change at different times, and different indicator data may also reflect the same information. Repeated emphasis on the same information can generate redundancy and affect the evaluation results, reducing the dimensionality of indicator data based on indicator sensitivity can avoid problems such as difficulty in determining the economic meaning of principal components and non unique factor loading matrices compared to commonly used principal component analysis methods [24]. Compare the evaluation results of indicators after dimensionality reduction optimization with those without dimensionality reduction treatment, as shown in Figure 4.1.

From Figure 4.1, it can be seen that there is a certain difference between the unit capability evaluation results after dimensionality reduction and the evaluation results without dimensionality reduction, but the difference is not significant. The reason for the certain difference is that after dimensionality reduction of the indicator data, some redundant information is reduced, avoiding the same information from being repeatedly emphasized. The evaluation results of the two are basically consistent, which can prove that the author's dimensionality reduction processing of indicator data based on indicator sensitivity is reasonable and effective. The reduced data information can represent the overall data information and accurately evaluate manufacturing capabilities.

Table 4.3: Weight values of various indicator combinations

Criterion layer	Indicator layer	Subjective power	Objective rights	C-B combination weight	C-A combination weight
Processing quality Q	Processing qualification rate	0.0702	0.553	0.628	0.1186
	Shape machining accuracy	0.0582	0.255	0.25	0.0453
	Dimensional machining accuracy	0.0413	0.196	0.135	0.0245
Processing cost C	Raw material consumption cost	0.0745	0.556	0.645	0.1267
	Labor management fee	0.0515	0.446	0.355	0.0702
	Arrival of raw materials	0.0607	0.238	0.295	0.0438
Manufacturing flexible F	Equipment utilization rate	0.0476	0.18	0.186	0.0275
	Equipment failure handling	0.0443	0.574	0.53	0.0775
	response time	0.0575	0.303	0.294	0.0533
Manufacturing time T	Processing time	0.0658	0.434	0.483	0.0868
	Auxiliary processing time	0.0499	0.266	0.226	0.0405
Environmental Protection E	Solid waste pollution	0.0536	0.446	0.508	0.0727
	Waste gas pollution	0.0447	0.32	0.297	0.0426
	Waste liquid pollution	0.0368	0.246	0.192	0.0276
Human reliability	Assignment difficulty	0.0535	0.165	0.187	0.0267
	Homework guidance	0.0447	0.192	0.185	0.0268
	Work skills	0.0367	0.158	0.127	0.0178
	physiological function	0.0535	0.18	0.198	0.0279
	Assignment time	0.0447	0.316	0.305	0.0433

Table 4.4: Total Capacity Values of Each Unit

	u_1	u_2	u_3	u_4	u_5	u_6
Processing quality	0.567	0.705	0.574	0.668	0.565	0.588
Processing cost	0.64	0.546	0.585	0.66	0.694	0.648
Manufacturing flexibility	0.558	0.658	0.677	0.64	0.689	0.558
Manufacturing time	0.534	0.667	0.605	0.594	0.588	0.68
environmental protection	0.465	0.458	0.585	0.557	0.420	0.633
Human reliability	0.593	0.623	0.575	0.67	0.567	0.574
Manufacturing capability value	0.563	0.586	0.58	0.607	0.570	0.645

At the same time, the author uses the sensitivity weight of the indicators to modify the indicator weight obtained by the entropy weight method, which can solve the problem of existing weighting methods only considering the fluctuation of the indicator's own data and ignoring the importance of the evaluation indicator to the entire evaluated object. After weighting the indicators, considering the temporal dynamics of unit indicator values, the time factor combined with grey correlation analysis method is finally introduced to obtain the capacity values of each unit. Compare the evaluation results of the method adopted by the author with those obtained from three methods: fuzzy analytic hierarchy process, rough set theory, and dynamic comprehensive evaluation, as shown in Figure 4.2.

From Figure 4.2, it can be seen that among the four methods, the unit capacity values obtained through dynamic horizontal and vertical comprehensive evaluation are most consistent with the capacity values obtained by the author's method. The capacity values obtained under rough set theory and fuzzy analytic hierarchy process have significant differences compared to these two methods, this is because rough set theory is a static evaluation method, while the method adopted by the author considers the temporal nature of the evaluation data. At the same time, fuzzy analytic hierarchy process is a subjective evaluation method, which has strong subjectivity and less attention to objective weights. Under the four methods, although there are differences in the capacity values of each unit, the fluctuation is within a reasonable range, indicating that the author's evaluation method is reasonable and feasible.

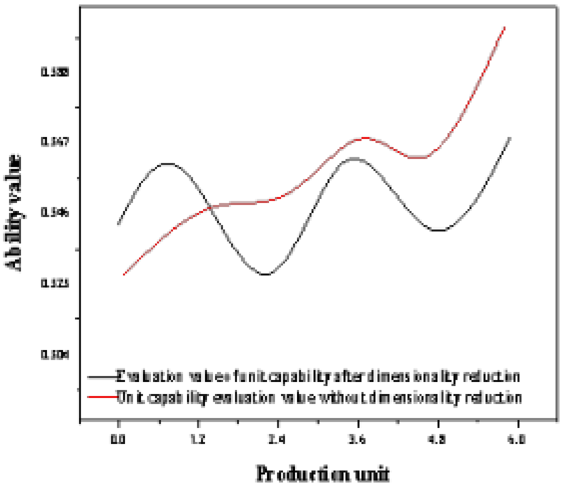


Fig. 4.1: Evaluation results of total manufacturing capacity of units before and after dimensionality reduction

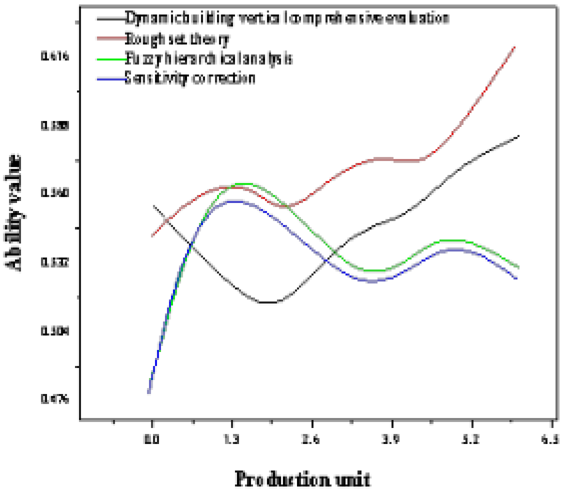


Fig. 4.2: Comparison of total manufacturing capacity values of units under different evaluation methods

In the evaluation process of unit manufacturing capability, the author analyzed the behavioral factors that affect human behavior in the manufacturing process, taking into account the role of human factors in the unit manufacturing process. By analyzing and evaluating the formation factors of each human factor behavior, the size of the unit's human reliability was obtained, and the unit manufacturing capability value was finally obtained by integrating factors such as unit time and quality. As shown in Figure 4.3, the figure reflects the ability values of each unit in processing quality, manufacturing flexibility, and other aspects. It can be seen from the figure that for production units with high human reliability, their ability in processing quality and

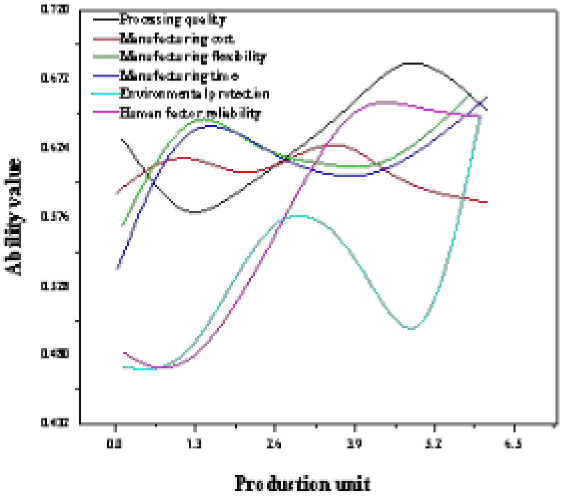


Fig. 4.3: Capability values of each unit in various aspects

manufacturing time is often higher, and the trend of change in the three is generally the same. In production and manufacturing, unreliable human behavior often leads to human error. Once a person makes an error, it often reduces their work efficiency and may also lower the production quality of the product to a certain extent. That is to say, human factors also have a certain impact on the unit manufacturing capacity. In the evaluation process of manufacturing capacity, the human factors in the production process cannot be ignored in order to obtain more accurate manufacturing capacity.

Evaluating the manufacturing capacity of production units can obtain the capability values of each unit in terms of processing quality, manufacturing flexibility, and other aspects. Therefore, workshop managers can timely grasp the production information of units, respond to weak links in unit production in a timely manner, and achieve optimal scheduling and complete production tasks on time.

5. Conclusion. The author established a manufacturing capacity evaluation model for production units. In response to the dynamic nature of data in the operation process of manufacturing production units and the varying importance of indicators to evaluation objects at different times, a combination of indicator sensitivity and entropy weighting method is proposed to objectively weight indicators, solving the problem of existing weighting methods only considering the fluctuation of indicator data and ignoring the importance of evaluation indicators to all evaluated objects. At the same time, in the process of unit manufacturing, humans are also an important component of the production unit, which can have a certain impact on the manufacturing capacity of the unit. Therefore, human reliability issues were considered in the unit capacity evaluation. Finally, the grey relational analysis method was used to obtain the various capabilities and comprehensive manufacturing capabilities of the unit at different times. Time dimension factors were introduced for time series data to obtain the total manufacturing value capabilities of each unit. Finally, the feasibility and effectiveness of the proposed evaluation method were verified through case analysis.

REFERENCES

[1] Baicun, W. A. N. G., Yuan, X. U. E., Jianlin, Y. A. N., **aoying, Y., & Yuan, Z. (2020). Human-centered intelligent manufacturing: overview and perspectives. Strategic Study of CAE, 22(4), 139-146.

[2] Bechinie, C., Zafari, S., Kroeninger, L., Puthenkalam, J., & Tscheligi, M. (2024). Toward human-centered intelligent assistance system in manufacturing: challenges and potentials for operator 5.0. Procedia Computer Science, 232, 1584-1596.

[3] Rožanec, J. M., Novalija, I., Zajec, P., Kenda, K., Tavakoli Ghinani, H., Suh, S., ... & Soldatos, J. (2023). Human-centric artificial intelligence architecture for industry 5.0 applications. International journal of production research, 61(20), 6847-6872.

[4] Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., ... & Xu, W. (2023). Six human-centered artificial intelligence grand challenges. International Journal of Human–Computer Interaction, 39(3), 391-437.

[5] Coronado, E., Kiyokawa, T., Ricardez, G. A. G., Ramirez-Alpizar, I. G., Venture, G., & Yamanobe, N. (2022). Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. Journal of Manufacturing Systems, 63, 392-410.

[6] Herrmann, T., & Pfeiffer, S. (2023). Kee** the organization in the loop: a socio-technical extension of human-centered artificial intelligence. AI & SOCIETY, 38(4), 1523-1542.

[7] Rannertshauser, P., Kessler, M., & Arlinghaus, J. C. (2022). Human-centricity in the design of production planning and control systems: A first approach towards Industry 5.0. IFAC-PapersOnLine, 55(10), 2641-2646.

[8] Froschauer, R., Kurschl, W., Wolfartsberger, J., Pimminger, S., Lindorfer, R., & Blattner, J. (2021). A human-centered assembly workplace for industry: Challenges and lessons learned. Procedia Computer Science, 180, 290-300.

[9] Romero, D., & Stahre, J. (2021). Towards the resilient operator 5.0: The future of work in smart resilient manufacturing systems. Procedia cirp, 104, 1089-1094.

[10] Selvaraj, V., & Min, S. (2023). Ai-assisted monitoring of human-centered assembly: A comprehensive review. International Journal of Precision Engineering and Manufacturing-Smart Technology, 1(2), 201-218.

[11] Zhang, R., Tan, H., & Afzal, W. (2021). A modified human reliability analysis method for the estimation of human error probability in the offloading operations at oil terminals. Process Safety Progress, 40(3), 84-92.

[12] Park, J., Jung, W., & Kim, J. (2020). Inter-relationships between performance sha** factors for human reliability analysis of nuclear power plants. Nuclear Engineering and Technology, 52(1), 87-100.

[13] Tang, B., & Cuschieri, A. (2020). Objective assessment of surgical operative performance by observational clinical human reliability analysis (OCHRA): a systematic review. Surgical endoscopy, 34, 1492-1508.

[14] Lin, C., Xu, Q. F., & Huang, Y. F. (2022). An HFM-CREAM model for the assessment of human reliability and quantification. Quality and Reliability Engineering International, 38(5), 2372-2387.

[15] Hardie, J. A., Green, G., Bor, R., & Brennan, P. A. (2021). Cutting edge selection: learning from high reliability organisations for virtual recruitment in surgery during the COVID-19 pandemic. The Annals of The Royal College of Surgeons of England, 103(6), 385-389.

[16] Wang, D., Wei, Y., Zhan, J., Xu, L., & Lin, Q. (2020). Human reliability assessment of home-based rehabilitation. IEEE Transactions on Reliability, 70(4), 1310-1320.

[17] Piechnicki, F., Dos Santos, C. F., De Freitas Rocha Loures, E., & Dos Santos, E. A. P. (2021). Data fusion framework for decision-making support in reliability-centered maintenance. Journal of Industrial and Production Engineering, 38(1), 1-17.

[18] Bednarek, M., & Dąbrowski, T. (2020). Selected tools increasing human reliability in the antropotechnical system. Journal of KONBiN, 50(2), 243-264.

[19] Appelganc, K., Rieger, T., Roesler, E., & Manzey, D. (2022). How much reliability is enough? A context-specific view on human interaction with (artificial) agents from different perspectives. Journal of Cognitive Engineering and Decision Making, 16(4), 207-221.

[20] Niu, H., Wu, W., **ng, Z., Wang, X., & Zhang, T. (2023). A novel multi-tasks chain scheduling algorithm based on capacity prediction to solve AGV dispatching problem in an intelligent manufacturing system. Journal of Manufacturing Systems, 68, 130-144.

[21] Song, M., Yang, M. X., Zeng, K. J., & Feng, W. (2020). Green knowledge sharing, stakeholder pressure, absorptive capacity, and green innovation: Evidence from Chinese manufacturing firms. Business Strategy and the Environment, 29(3), 1517-1531.

[22] Ghani, N. F. A., Zaini, S. N. A. M., & Abu, M. Y. (2020). Assessment the unused capacity using time driven activity based costing in automotive manufacturing industry. Journal of Modern Manufacturing Systems and Technology, 4(1), 82-94.

[23] Hamidu, Z., Boachie-Mensah, F. O., & Issau, K. (2023). Supply chain resilience and performance of manufacturing firms: role of supply chain disruption. Journal of Manufacturing Technology Management, 34(3), 361-382.

[24] Hemalatha, C., Sankaranarayanasamy, K., & Durairaaj, N. (2021). Lean and agile manufacturing for work-in-process (WIP) control. Materials Today: Proceedings, 46, 10334-10338.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 11, 2024
Accepted: Jul 20, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1257–1266, DOI 10.12694/scpe.v26i3.4131

OPTIMIZATION METHOD OF CALIBRATION CYCLE BASED ON STATE EVALUATION RESULTS OF ELECTRIC ENERGY METERS

YING ZHANG[*] WENJING WANG[†] SHI CHEN [‡] ZILIN CHEN [§] SHU CAO [¶] AND YINTING GUO[‖]

Abstract. In order to solve the problems of heavy workload, weak planning, and repetitive maintenance in the periodic rotation of smart energy meters, the author proposes a verification cycle optimization method based on the evaluation results of energy meter status. This method first obtains data on six indicators of smart energy meters: regional factors, reliability, full event, abnormal metering events, battery overload, and clock battery undervoltage; Subsequently, on the one hand, the coefficient of variation assignment method is used to obtain the status score of each electricity meter, and on the other hand, these six indicator data are used as input data, and the K means clustering algorithm is used to classify and obtain the corresponding categories. Finally, the two algorithms are combined to obtain a new method for evaluating the status of smart energy meters, and the final evaluation result is output. The experimental results indicate that: The number of electricity meters scored below 80 points obtained by this method accounts for 22.08% of the total number of electricity meters, while electricity meters scored above 80 points account for 77.93% of the total number of electricity meters. This indicates that this method is in line with the actual situation and objective laws. Constructing a state evaluation model for electric energy meters, using historical data and on-site calibration data as state variables, analyzing the annual operational quality of electric energy meters, and providing reference basis for adjusting the calibration cycle of electric energy meters.

Key words: State evaluation methods, K means, Coefficient of variation method, Electricity meter, Error, Verification cycle

1. Introduction.

Electricity meters are a bridge between power supply enterprises and electricity customers for billing and settlement, and an important measuring tool for people's daily electricity consumption [1]. With the rapid development of smart grids, smart energy meters are widely used due to their powerful functions, high measurement accuracy and sensitivity. The quality of their operation directly affects the economic benefits of power grid enterprises and the vital interests of users [2]. With the application of new technologies and methods, the level of power management is also constantly improving. Improving the lean management level of electricity metering and rational allocation of human, financial and resources have become important needs in the new era. The following uses big data analysis methods to explore and study the optimization of the re inspection cycle of electricity meters [3]. Mining data on electricity meter calibration and resource loss, based on discrete degree analysis and EUAC (equivalent comprehensive cost) analysis method, provides reference for optimizing calibration cycle, improving management level, assisting departmental decision-making, and provides ideas for wider applications [4].

Create a state evaluation model to evaluate the status of electricity meters. This model is based on the current and historical data of the electricity meter, and applies membership function to establish and solve fuzzy technology to design a state quantity evaluation model. It combines entropy weight method to objectively evaluate the operation quality of the electricity meter [5]. The state evaluation model includes the selection of state variables, normalization of state variables, evaluation of state components, and overall state evaluation of smart energy meters.

With the improvement of the production level of smart energy meters, the drawbacks of the traditional one size fits all disassembly and calibration method for periodic calibration of energy meters have become increasingly apparent [6]. In order to solve the problem of dismantling and re calibrating smart energy meters

[*] State Grid Fujian Marketing Service Center, Fuzhou, Fujian, 350001, China.(Corresponding author, YingZhang731@126.com)
[†] State Grid Fujian Marketing Service Center, Fuzhou, Fujian, 350001, China.(WenjingWang9@163.com)
[‡] State Grid Fujian Marketing Service Center, Fuzhou, Fujian, 350001, China.(ShiChen833@126.com)
[§] State Grid Fujian Marketing Service Center, Fuzhou, Fujian, 350001, China.(ZilinChen7@163.com)
[¶] State Grid Fujian Marketing Service Center, Fuzhou, Fujian, 350001, China.(ShuCao7@126.com)
[‖] State Grid Fujian Marketing Service Center, Fuzhou, Fujian, 350001, China.(YintingGuo9@163.com)

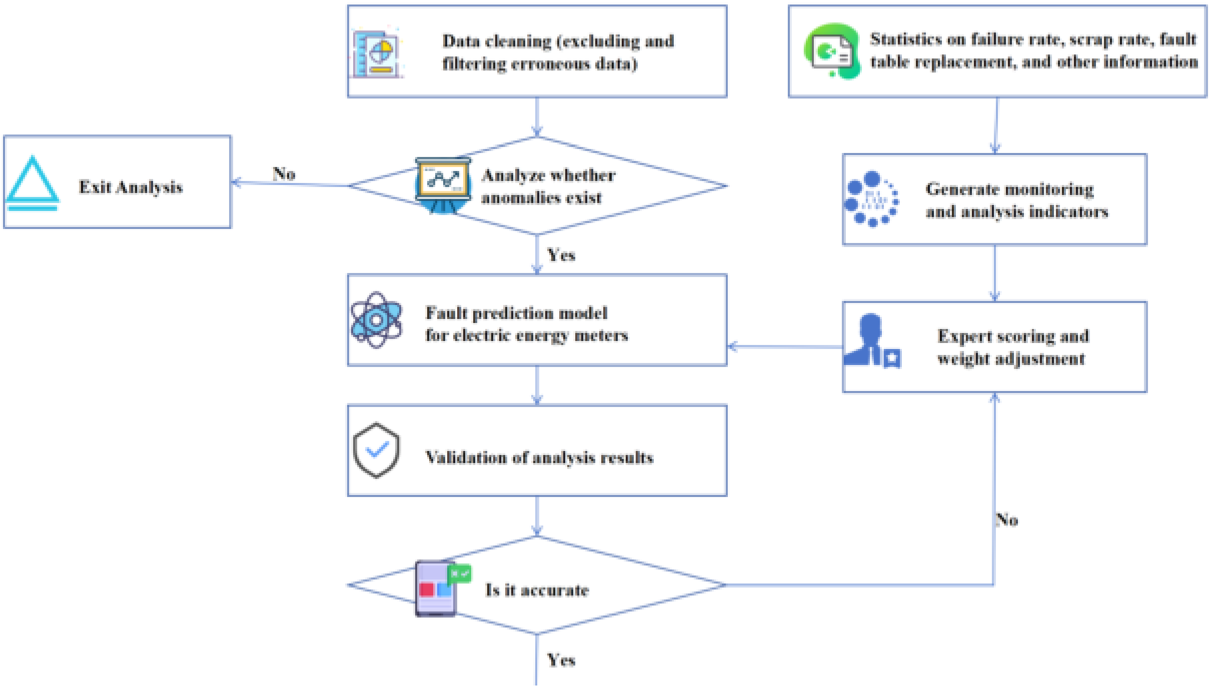

Data cleaning (excluding and filtering erroneous data)
Statistics on failure rate, scrap rate, fault table replacement, and other information
Exit Analysis
No
Analyze whether anomalies exist
Generate monitoring and analysis indicators
Yes
Fault prediction model for electric energy meters
Expert scoring and weight adjustment
Validation of analysis results
No
Is it accurate
Yes

Fig. 1.1: Energy meter status evaluation

within an eight year calibration cycle, the author conducted a research on quality supervision and evaluation of energy meters based on full life cycle testing data. Based on the first inspection data of the electricity meter, big data analysis technology is used to assess the risk of the electricity meter; Conduct periodic verification through on-site verification based on risk screening results; Analyze the calibration error of the electric energy meter based on the first inspection and on-site verification data; Establish a state evaluation model to evaluate the operational quality of electric energy meters, providing a reference basis for extending the cycle of electric energy meters. As shown in Figure 1.1.

2. Literature Review. As a measuring and recording device for power supply and consumption, the operation status of smart energy meters not only directly affects the development and operational efficiency of power grid companies, but also affects the fairness and safety of user electricity consumption [7]. At present, smart energy meters have been widely used in important links such as power generation, transmission, distribution, and consumption. As a key component of the smart grid, smart energy meters play an important role in real-time measurement, load control, and response to power demand in power grid companies [8]. The power grid company vigorously promotes the comprehensive construction of big data technology in the smart grid, and various emerging technologies are widely applied in the power industry. Among them, the establishment of databases such as substation data, distribution data, electricity consumption data, and marketing data is also becoming increasingly perfect, providing a good research environment for the research and development of big data technology [9].

At present, Power Grid Co., Ltd. has completed the construction and operation of the corresponding data management platform, gradually leveraging the role of big data technology in electricity data management and analysis [10]. However, current data analysis techniques cannot fully utilize smart energy meter data to achieve ideal analysis results, and continuous research and practice are still needed. Therefore, using big data technology to analyze electricity meter data is the development trend of future smart grid technology, and the increasingly perfect big data platform and corresponding technology make it have great potential for development. Pazderin,

A. V. used a state estimation method based on the direct measurement data of EE in watt hours (volt ampere reactive hours) provided by an electricity meter to determine the EE flow rate. EFP solutions are essential for a wide range of applications, including instrument data validation, zero imbalance EE billing, and non-technical EE loss checks [11]. Singh, M. provides a detailed description of some of the challenges faced by electricity consumption data, including saving large amounts of data, deleting, manipulating, and adjusting data. Blockchain is a promising technology that can use encryption algorithms to address issues of data integrity and confidentiality [12]. Dakyen, M. M. et al. used big data technology to analyze the electricity consumption data of smart energy meters, in order to better evaluate the status of smart energy meters [13]. Jie YANG has modularized various data, evaluation index systems, and evaluation methods for the current state evaluation indicators of electric energy metering devices, aiming to address the uncertainty of the evaluation indicators and the inconsistency of the results of various evaluation methods. He has dynamically built an electric energy metering device state evaluation system, which provides a comprehensive description of the state indicators, but does not involve the detailed differences in the state indicators of each component of the electric energy metering device [14]. The multi-objective comprehensive evaluation method for smart meter suppliers based on grey correlation degree described in Zhu, X, while retaining the advantages of the multi-objective comprehensive evaluation model, solves the problems of cumbersome indicators and strong subjectivity in traditional smart meter supplier comprehensive evaluation [15]. But this method only analyzes the overall batch of electricity meters, ignoring the evaluation of individual electricity meter states.

A new method for evaluating the status of smart energy meters is proposed by combining the coefficient of variation assignment method and K-means clustering algorithm. The evaluation results of smart energy meters are obtained by analyzing six indicators: regional impact index, all events, measurement anomalies, meter overload, and clock battery undervoltage. The rationality of the method is analyzed based on the results.

3. Method.

3.1. Design Concept. Firstly, based on the previous statistical analysis and research, six indicators that have a significant impact on the evaluation of the status of smart energy meters were identified, namely meter reliability, regional factors, all events, measurement anomalies, meter overload, and clock battery undervoltage. Then, these indicators were analyzed using two methods: One was weighted using the coefficient of variation method, by analyzing the impact of different indicators on the operating status of electricity meters, assign corresponding weights to each indicator, and finally reflect the operating status of each electricity meter in the form of a score; Another approach is to use these indicators as features of the electricity meter, treating the meter as points in space, where these indicators are the coordinates of the points. Clustering methods are used to classify these points and obtain different evaluation states of the electricity meter [16]. Combine the evaluation results of these two methods to obtain the final evaluation result of the operating status of the electricity meter. The process is shown in Figure 3.1.

3.2. Indicator data.

1) Reliability of electric energy meters. The reliability calculation formula for electric energy meters is:

$$M_r = 1 - \frac{\sum_1^t f(1)}{N} \tag{3.1}$$

In the formula: M_r is the reliability index of the electric energy meter; $f(i)$ is the number of faulty meters in the i-th month of the current batch of electricity meters; N is the total number of electricity meters in the current batch; t is the current month.

2) Regional factors. The formula for calculating regional factor indicators is:

$$M_t = 1 + \frac{H_x}{H} \times lg(\frac{H_x}{H} \cdot \frac{J}{J_x}) \tag{3.2}$$

In the formula: M_t is the regional factor indicator; H_x is the number of electricity meters installed in the x-th city; J_x is the total number of installed energy meters; J is the number of faulty energy meters in the x-th city; is the total number of faulty energy meters.

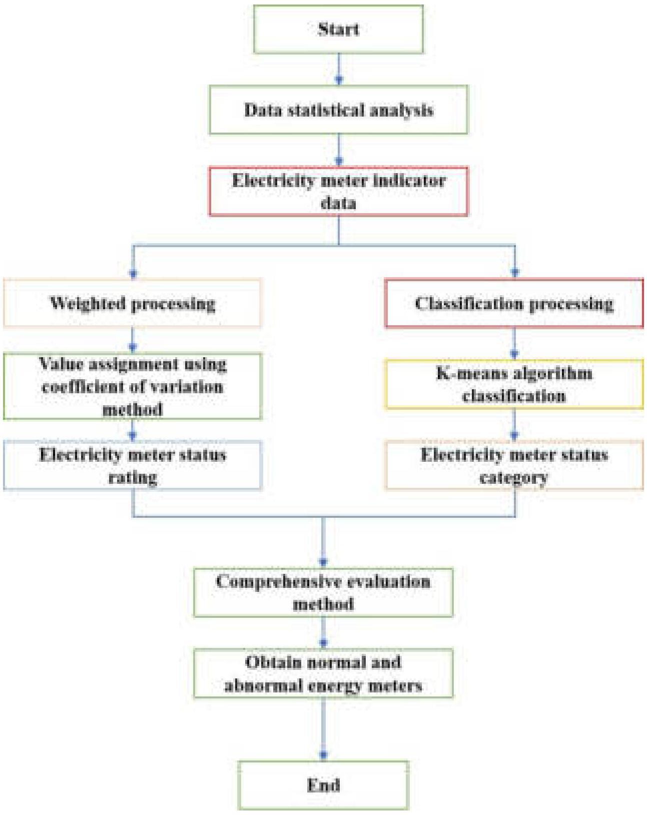

Start
Data statistical analysis
Electricity meter indicator data
Weighted processing
Classification processing
Value assignment using coefficient of variation method
K-means algorithm classification
Electricity meter status rating
Electricity meter status category
Comprehensive evaluation method
Obtain normal and abnormal energy meters
End

Fig. 3.1: Design process of state evaluation method

3) Full event. The calculation formula for all event indicators is:

$$M_q = \begin{cases} 100 \times \sum_{i=1}^{11} P(A_2|C_i), \sum_{i=1}^{11} P(A_2|C_i \leqslant 1) \\ 100, \sum_{i=1}^{11} P(A_2|C_i > 1) \end{cases} \tag{3.3}$$

In the formula, M_q represents the overall event indicator; A_2 indicates that the calibration result of the electric energy meter is faulty; C_i $(i = 1, 2, \cdots, 11)$ refers to 11 events, including meter shutdown, meter runaway, meter deviation, meter reverse connection, meter time deviation, meter power outage frequency change, meter phase failure frequency, magnetic field anomaly, meter transformer rate change, cover opening recording, and time synchronization; $P(A_2|C_i)$ represents the probability of C_i occurring under A_2 conditions.

4) Abnormal measurement. The formula for calculating abnormal measurement indicators is:

$$M_a = \begin{cases} 100 \times (\sum_{i=1}^{6} P(A_2|B_j) + \sum_{k=1}^{5} y_k), \sum_{i=1}^{6} P(A_2|B_j) + \sum_{k=1}^{5} y_k \leqslant 1 \\ 100, \sum_{i=1}^{6} P(A_2|B_j) + \sum_{k=1}^{5} y_k > 1 \end{cases} \tag{3.4}$$

In the formula, M_a represents the measurement anomaly indicator; $B_j (j = 1, 2, \cdots, 6)$ refers to uneven energy representation, meter flying, meter reversing, meter stopping, abnormal reverse power and clock; $y_k (k = 1, 2, \cdots, 5)$ represents the correlation between voltage exceeding limits, voltage loss, current overcurrent, voltage disconnection, and reverse flow events and anomalies; $P(A_2|B_j)$ represents the probability of A_2 occurring under B_j conditions.

5) The electricity meter is overloaded. The formula for calculating the overload index of an electric energy meter is:

$$M_1 = \begin{cases} 100K_w \times log_2(\frac{W_0}{W_N}), K_W > 0 \\ 0, K_W = 0 \end{cases} \tag{3.5}$$

In the formula: M_1 is the overload indicator of the electric energy meter; W_N is the amount of electricity measured within 24 hours of normal rated operation of the energy meter; K_W is the proportion of days in which the daily electricity consumption exceeds the standard metering electricity of the electricity meter within 6 months; W_0 is the average daily electricity consumption of the portion of electricity consumption that exceeds the standard measurement of the electricity meter within 6 months.

6) Clock battery undervoltage. The formula for calculating the undervoltage index of the clock battery is:

$$M_c = \frac{100 - 100e^{-x}}{1 + e^{-z}} \tag{3.6}$$

In the formula, M_c represents the undervoltage indicator of the clock battery; Z is the number of clock undervoltages that occur within 6 months.

3.3. Principle of coefficient of variation method. The coefficient of variation assignment method directly utilizes the information contained in various indicators to calculate the weights of the indicators, and is an objective weighting method [17]. Based on the impact of changes in indicator data on the evaluation results of electricity meters, further analyze the importance of this indicator in the evaluation of results [18]. Reflected in numerical terms, the greater the degree of variation of the indicator data, the greater the assigned value to the indicator.

The coefficient of variation chosen by the author is the standard deviation, and its main calculation steps are as follows.

1). Assuming there are m objects to be evaluated and a total of n evaluation indicators, the evaluation matrix X of the indicators can be expressed as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3.7}$$

In the formula, x_{ab} represents the characteristic data of the $a(a = 1, 2, \cdots, n)$ evaluation indicator for the b (b=1,2, m) th evaluation object.

2). Calculate the average value $\overline{x_a}$ of the a-th indicator as follows:

$$\overline{x_a} = \frac{1}{m} \sum_{b=1}^{m} x_{ab} \tag{3.8}$$

3). Calculate the standard deviation σ_a of the a-th indicator as:

$$\sigma_a = \sqrt{\frac{1}{m} \sum_{b=1}^{m} (x_{ab} - \overline{x_a})^2} \tag{3.9}$$

4). Calculate the coefficient of variation for the a-th indicator as:

$$v_a = \frac{\sigma_a}{\overline{x_a}} \tag{3.10}$$

5). Normalize the obtained coefficient of variation and calculate the objective weight β_a of the a-th indicator as follows:

$$\beta_a = \frac{v_a}{\sum_{a=1}^{n} v_a} \tag{3.11}$$

3.4. K-means clustering principle. The core idea of the K means algorithm is to randomly select data from the dataset as the initial clustering center, and calculate the distance from other data points to these data points. These data points are divided into the closest clustering centers. After traversing all the data, the average value of each class of data is used as the new clustering center, and the above operation is repeated again until a certain threshold is met or a predetermined number of iterations are reached before stopping [19].

The specific steps of the K-means algorithm are mainly divided into the following steps:

1. Based on a known dataset, use k data points as the initial cluster center C, where these k data points are arbitrarily selected;
2. Calculate the Euclidean distance between data samples other than the cluster center and the cluster center;
3. Using Euclidean distance as the basis for judgment, divide the data samples into clusters belonging to the cluster center closest to them;
4. Calculate the mean of the data samples in each cluster and use it as the new cluster center for each cluster to calculate the sum of squared errors for this dataset;
5. Determine whether the total sum of squared errors of the entire dataset remains unchanged or fluctuates within a small range. If so, end the clustering and output the final clustering result; Otherwise, go back to step 2) and loop in the order of steps until the requirements are met or the set number of iterations is reached.

In practical calculations, all events, measurement anomalies, meter overload, and clock battery undervoltage are converted into indicator data of 1-100 using equations 3.3-3.6 to achieve unity of magnitude, and then clustering algorithm calculations are performed [20]. The Euclidean distance formula is:

$$D(z, E_p) = \sqrt{\sum_{q=1}^{Q} (z_q - E_{pq})^2} \tag{3.12}$$

In the formula: z is the data sample; E_p is the p-th cluster center; Q is the dimension of the data sample; z_q, E_{pq} is the qth feature of z and E_p.

4. Results and Discussion. Select the operating data of all dismantled and calibrated electricity meters in a certain area for the 6 months before the evaluation time, and analyze and process these data to obtain the indicator data for the evaluation of smart electricity meters, this includes the reliability of smart energy meters, regional factors, all events, metering anomalies, meter overload, and clock battery undervoltage. The reliability indicators and regional factors are obtained based on the data analysis of the entire province's electricity meters from installation to evaluation time, while the weights of each indicator are obtained using the coefficient of variation assignment method based on the average indicator data of each batch of electricity meters in each month throughout the year, including all events, measurement anomalies, meter overload, and clock battery undervoltage. And based on the data from the 6 months before the evaluation time, obtain the indicator data of the current status of each smart energy meter, and finally calculate the current operating status of each smart energy meter by combining reliability indicators and regional factors.

The state evaluation of smart energy meters is defined as:

$$R = (100 - w_1 M_q - w_2 M_q - w_3 M_q - w_4 M_q) \times M_r \times M_t \tag{4.1}$$

In the formula, R represents the evaluation result of each energy meter; $w_1 - w_4$ is its corresponding weight, obtained by the coefficient of variation assignment method.

The experimental case analysis is based on data from 12 batches of electricity meters evaluated in July 2021, each batch including disassembled and still running electricity meters. The weight of the indicators is calculated

Table 4.1: Energy meter status evaluation Table

Score	Number of electricity meters/piece	Proportion/%	Predicted number of faults/block	Predicted proportion of faults/%	Normal quantity/block in the predicted number of faults
0∼60	4250	5.15	808	29.22	0
60∼70	11050	13.34	877	31.67	0
70∼80	2967	3.59	342	12.34	0
80∼90	12867	15.55	526	18.94	19
90∼100	51587	62.37	235	7.83	198
amount	69854	100.00	2788	100.00	217

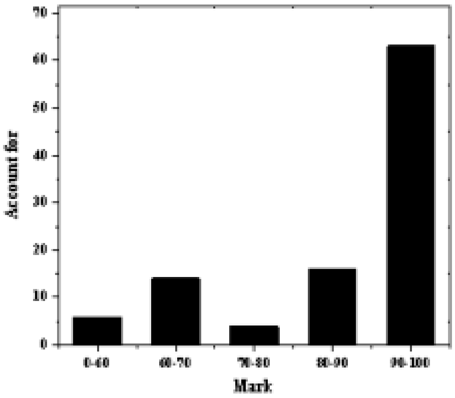


Fig. 4.1: Distribution of Energy Meter State Evaluation

based on the average monthly indicator data of the batch of electricity meters in 2021. As the dismantling of smart electricity meters requires a certain process and time, the quality of the evaluation results is evaluated based on the dismantling of the meters within 3 months after the evaluation results. The evaluation results are shown in Table 3.1, Figure 4.1, and Figure 4.2.

Table 4.1 shows the specific quantity of each score segment, where the number of faults is based on the number of fault tables obtained through disassembly and detection within three months after the current evaluation time point, and the normal number is based on the number of normal tables obtained through disassembly and detection within three months after the current evaluation time point. Figures 3.2 and 3.3 are visualizations of the distribution of all meter quantities and the distribution of fault meter quantities for the evaluation of the energy meter status in Table 4.1, respectively.

From Table 4.1 and Figure 4.1, it can be seen that 22.08% of the total number of electricity meters are scored below 80 points according to this method, and 77.93% are scored above 80 points. At the same time, it can be seen from Table 4.1 and Figure 4.2 that the number of faults in electricity meters scored below 80 points accounts for 73.26% of the total number of faulty electricity meters. This indicates that the method is in line with the actual situation and objective laws.

4.1. K-means algorithm analysis. This evaluation method considers smart energy meters as points in space, and considers the reliability, regional factors, all events, measurement anomalies, meter overload, and clock battery undervoltage as the coordinates of points in the space. These coordinates are used as inputs to the K-means algorithm, and points with similar distances can be clustered in the same area based on the distance between points to achieve classification results.

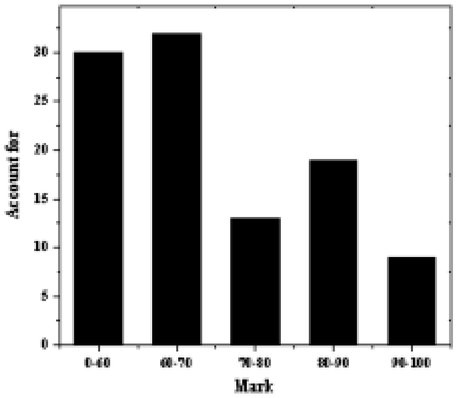


Fig. 4.2: Distribution of the number of faulty energy meters

Table 4.2: Energy meter status evaluation Table

Score	Number of electricity meters/piece	Proportion/%	number of faults/block	Fault proportion of faults/%	Normal quantity/block in the predicted number of faults
Class I	63905	77.26	798	28.18	215
Class II	207	0.26	12	0.35	0
Class III	867	1.06	62	2.23	1
Class IV	2437	2.96	291	10.47	0
Class V	15303	18.46	1625	58.77	0
total	82719	100.00	2788	100.00	216

Using the data from the smart electricity meter batch mentioned above for analysis, referring to the analysis results of the coefficient of variation method, based on data characteristics and the principle of facilitating result comparison and analysis, this method determines that the K-means clustering algorithm has 5 categories. The clustering results are shown in Table4.2, Figure 4.3, and Figure 4.4.

Table 4.2 shows the specific quantity of each score segment, where the number of faults is based on the number of fault Tables detected by dismantling within 3 months after the current evaluation time point, and the normal number is based on the number of normal Tables detected by dismantling within 3 months after the current evaluation time point. Figures 4.4 and 4.5 visualize the distribution of the number of all meters and the distribution of the number of faulty meters in the state classification of electricity meters in Table 4.2, respectively.

From Table 4.2 and Figure 4.5, it can be seen that Class I electricity meters have the most data, as the analysis includes both disassembled and still running smart electricity meters, so normal electricity meters account for the majority. Obviously, Class I should be considered as a normal energy meter category, while Class II-V should be considered as an abnormal energy meter category. The classification here is for comparison with the coefficient of variation method. According to the K-means clustering algorithm, 77.26% of energy meters are classified as normal energy meters, and 71.82% of actual faulty energy meters are included in the category of abnormal energy meters. The abnormal energy meters analyzed in the K-means algorithm's energy meter status evaluation include most of the actual faulty energy meters, which is consistent with the actual situation, indicating that the evaluation result has a certain degree of scientific and rationality.

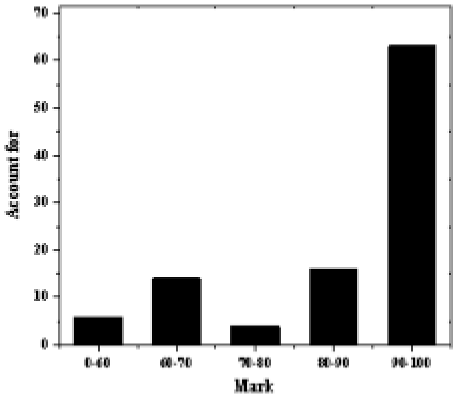


Fig. 4.3: K-means clustering distribution of all electricity meters

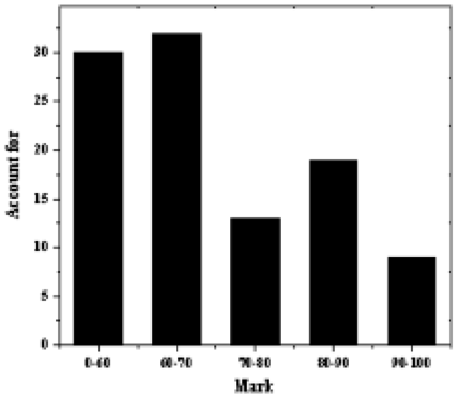


Fig. 4.4: K-means clustering distribution of fault Table

5. Conclusion. The author proposes a study on the optimization method of calibration cycle based on the evaluation results of electricity meter status, introduces the principles of coefficient of variation algorithm and K-means algorithm, and evaluates the status of electricity meters based on these two algorithms. After comparing the evaluation results of coefficient of variation method and K-means algorithm, combined with the characteristics of electricity meter evaluation parameters, a new state evaluation method is constructed by integrating the coefficient of variation method and K-means algorithm, prove its scientific and feasibility through data analysis, providing new ideas for the state evaluation of smart meters. This method has been recognized by the power supply company in practical experiments.

6. Acknowledgement. State Grid Corporation of China Headquarters Management Technology Project (5700-202227226A-1-1-ZN): Research on Key Technologies for Informationization Evaluation and Calibration of Household Energy Meters Based on Non convex Sparse Multi objective Optimization.

REFERENCES

[1] Yan, R., Zheng, Y., Yu, N., & Liang, C. (2024). Multi-smart meter data encryption scheme basedon distributed differential privacy. Big Data Mining and Analytics, 7(1), 131-141.

[2] Jiang, T., Shen, Z., Jin, X., Zhang, R., Parisio, A., & Li, X., et al. (2023). Solution to coordination of transmission and distribution for renewable energy integration into power grids:an integrated flexibility market. Journal of Electric Power and Energy Systems, Chinese Society of Electrical Engineering, 9(2), 444-458.

[3] Naeemah, A. J., & Wong, K. Y. (2023). Selection methods of lean management tools: a review. International Journal of Productivity and Performance Management, 72(4), 1077-1110.

[4] Jie YANG, Xiaoshu CAO, Jun YAO, Zhewen KANG, Jianxia CHANG, & Yimin WANG. (2024). Geographical big data and data mining: a new opportunity for "water-energy-food" nexus analysis. Journal of Geography (English Edition), 34(2), 203-228.

[5] Zhang, J., Li, B., Peng, Q., & Gu, P. (2023). Product specification analysis for modular product design using big sales data. Chinese Journal of Mechanical Engineering: English Edition, 36(1), 19-33.

[6] Liao, Y., Weng, Y., Tan, C. W., & Rajagopal, R. (2022). Quick line outage identification in urban distribution grids via smart meters. CSEE Journal of Power and Energy Systems, 8(4), 1074-1086.

[7] Kong, F., Yin, S., Sun, C., Yang, C., Chen, H., & Liu, H. (2022). Design and optimization of a maglev electromagnetic ribo-electric hybrid energy converter for supplying power to intelligent sensing equipment. Sustainable Energy & Fuels, 6(3), 800-814.

[8] Ahmed, M., Khan, A., Ahmed, M., Tahir, M., Jeon, G., & Fortino, G., et al. (2022). Energy theft detection in smart grids:taxonomy,comparative analysis,challenges,and future research directions. Journal of Automation: English Edition, 9(4), 23.

[9] vorah. (2022). A comparison on pso optimized pid contro11er for inter area osci11ation contro1 in an interconnected power system. Technology and Economics of Smart Grids and Sustainable Energy, 7(1), 1-14.

[10] Bhaskar, U., Mishra, B., Yadav, N., & Sinha, P. (2023). Who uses deceptive impression management to succeed at job interviews? the role of ethical ideologies and work locus of control. International Journal of Manpower, 44(3), 453-469.

[11] Pazderin, A. V., Polyakov, I. D., & Samoylenko, V. O. (2022). Energy flow problem solution based on state estimation approaches and smart meter data. Global Energy Internet: English, 5(5), 551-563.

[12] Singh, M., Ahmed, S., & Sharma, S. (2022). Blockchain-based smart electricity measurement and monitoring system: a survey. i-Manager s Journal on Embedded Systems, 11(1), 12-16.

[13] Dakyen, M. M., Dagbasi, M., & Zdenefe, M. (2022). Energy models for cost-optimal analysis: development and calibration of residential reference building models for northern cyprus:. Indoor and Built Environment, 31(3), 657-681.

[14] Li, X., Hu, Y., Xue, B., Wang, Y., Zhang, Z., & Li, L., et al. (2022). State-of-health estimation for the lithium-ion battery based on gradient boosting decision tree with autonomous selection of excellent features. International Journal of Energy Research, 46(2), 1756-1765.

[15] Zhu, X., Cui, J., & Li, Y. (2023). Energy savings bottleneck diagnosis of cooling system based on integrated correlation analysis. The Canadian Journal of Chemical Engineering, 101(8), 4762-4770.

[16] Wang, C., Qin, D., Wen, Q., Zhou, T., Sun, L., & Wang, Y. (2022). Adaptive probabilistic load forecasting for individual buildings. iEnergy, 1(3), 341-350.

[17] Meng, F., & Tian, K. (2022). Interval type-2 fuzzy logic based radar task priority assignment method for detecting hypersonic-glide vehicles. Frontiers of Information Technology & Electronic Engineering, 23(3), 488-501.

[18] Kok, K., Paterakis, N. G., & Giraldo, J. S. (2022). Development, application, and evaluation of an online competitive simulation game for teaching electricity markets. Computer Applications in Engineering Education, 30(3), 759-778.

[19] Bennaceur, H., Almutairy, M., & Alhussain, N. (2023). Genetic algorithm combined with the k-means algorithm:a hybrid technique for unsupervised feature selection. Intelligent Automation and Soft Computing, 37(9), 2687-2706.

[20] Bai, R., Shi, Y., Yue, M., & Du, X. (2023). Hybrid model based on k-means++ algorithm, optimal similar day approach, and long short-term memory neural network for short-term photovoltaic power prediction. Global Energy Internet: English, 6(2), 184-196.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 11, 2024
Accepted: Sep 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1267–1275, DOI 10.12694/scpe.v26i3.4200

THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN LOGISTICS MANAGEMENT OPTIMIZATION RESEARCH

CHUNXIANG WU [*]AND RONG WANG[†]

Abstract. In order to solve the problem of low efficiency and low accuracy in manual data collection in traditional logistics management systems and traditional warehouse management, the author proposes the application of artificial intelligence in logistics management optimization research. The author designed an RFID based warehouse management system that includes management layer, intermediate layer, and physical layer, focusing on the warehousing and warehousing operations in the production and manufacturing workshop. This system can optimize the traditional warehousing process, dynamically and intelligently perceive management objects, and achieve effective management and monitoring of warehousing management. Establish a simulation model using Flexsim software to simulate the inbound process and simulate the model entity in a relatively short period of time. The experimental results show that the total utilization rate of AGV in the simulation model reaches 72%. This method effectively solves the problems of slow data collection and low accuracy in traditional logistics warehousing processes.

Key words: Smart warehousing, Manufacturing workshop, Digital carrier, RFID, Inbound operations, Warehouse management system

1. Introduction. With the advent of the Internet era, e-commerce has an increasingly significant impact on people, greatly promoting the development of the logistics industry [1]. At present, the development level and scale of e-commerce have been in a leading position. With the in-depth implementation of the national ”Internet plus” action plan, the scale of e-commerce market has continued to expand, e-commerce service products have been widely popularized, and higher requirements have been put forward for the logistics industry [2]. Therefore, enterprises need to give full play to the advantages of artificial intelligence technology and use Internet, big data, cloud computing and other technologies to provide new ideas for logistics management [3,4]. The use of artificial intelligence technology by enterprises in the traditional logistics industry can effectively improve logistics management efficiency while providing consumers with higher quality and efficient services and products.

Artificial intelligence technology can promote the informatization and logistics of enterprises in a sense, and has great significance in improving the economic and social benefits of enterprises [5]. Taking the world’s express logistics as an example, due to various constraints, there are still some problems in the combination of world logistics and artificial intelligence technology, such as long delivery cycles, slow delivery speeds, and low delivery rates of express delivery [6]. In order to address these issues, express delivery companies have begun to utilize artificial intelligence technology. For example, an automated warehousing and logistics system based on AI technology has been established, making the entire process of operations more intelligent and efficient [7]. Artificial intelligence can be applied in multiple aspects of logistics management, from warehouse management to supply chain management, from logistics distribution systems to logistics planning systems, all of which can be optimized through artificial intelligence [8]. In practical applications, artificial intelligence can be effectively integrated with logistics technology to improve logistics management through advanced technologies such as machine learning. At present, the logistics industry is in a critical period of transformation and development, which requires the improvement of logistics management level through intelligent technology. Artificial intelligence is a high-tech means, and its advanced algorithms can effectively solve various problems that arise in the logistics management process, thereby improving the overall development level of the logistics industry [9,10].

[*] Qinhuangdao Vocational and Technical College, 066100, China.(ChunxiangWu2@126.com)

[†] Qinhuangdao Vocational and Technical College, 066100, China.(Corresponding author’s email:RongWang89@163.com)

2. Literature Review. With the development of artificial intelligence, logistics services need to be constantly updated and iterated to meet the service needs of enterprises at different stages and industries [11]. Based on artificial intelligence technology, logistics enterprises can integrate warehousing and distribution processes, effectively control each link in the transportation process, and thereby improve the transportation efficiency of goods [12]. Taking the express delivery industry as an example, it has now achieved automatic pickup function and is continuously promoting the development of unmanned vehicle delivery mode, which also puts higher requirements on delivery services. In this mode, the intelligent delivery robot achieves real-time positioning and navigation of goods through an onboard system; Avoiding the risk of obstacles during vehicle operation through autonomous path planning; At the same time, it will also engage in voice communication with users to understand their needs [13-14]. Today, with the rapid development of artificial intelligence, many industries have undergone tremendous changes. The introduction of artificial intelligence technology into the modern logistics industry has promoted the digital transformation of the logistics industry. As a major component of modern logistics industry, energy logistics has a broad operating scale and huge market capacity. Therefore, in the process of national development, energy logistics will inevitably be highly valued by people. Zhang, J. et al. proposed the Internet of Things intelligent logistics network. The integration of various applications in logistics systems into many systems is costly and in some cases not used, making it difficult to justify their purchase. IoE is important in information and communication technology because it can generate large amounts of data and use various mathematical techniques to evaluate the complex connections between the transactions represented by this data [15]. Yang, X. et al. proposed an optimization algorithm for logistics transportation costs of prefabricated building components suitable for project management. We constructed a project management oriented prefabricated building component management system, analyzed the logistics and transportation process of prefabricated building components, and scheduled logistics vehicles for prefabricated components within a time window [16]. Zhao, H. and others combined the characteristics of logistics distribution to mathematically design the distribution vehicle routing problem. Introduce the mountain climbing process with strong local search ability into the particle swarm optimization (PSO) process to improve the provided method [17].

The author introduces digital carrier technology and designs a warehouse management platform based on Radio Frequency Identification (RFID), which solves the problem of difficult and error prone data collection in traditional manufacturing workshop management processes. According to the principles of material identity, similarity, complementarity, and first in, first out, intelligent analysis of storage capacity is carried out to ensure that suitable goods are stored in appropriate locations and monitored. In addition, the author used Flexsim software to establish a production and manufacturing warehouse workshop model, analyzed simulation output results, identified bottlenecks in system design, improved warehouse operation efficiency, and saved operating costs.

3. Method.

3.1. Smart warehousing. Smart warehousing logistics is a crucial part of smart manufacturing, which combines the scattered personnel and cluttered information in traditional warehousing workshops to enable enterprises to operate smoothly and efficiently [18]. Smart warehousing logistics replaces rigid and high maintenance cost elevated warehouses with more flexible and lower maintenance cost building storage methods, replaces manned forklifts with more flexible and intelligent mobile robots, replaces traditional equipment that is prone to single point failures with more reliable equipment, and replaces complex and highly specialized traditional equipment with simpler structure, intelligent control, and easy maintenance intelligent equipment [19]. Traditional warehousing logistics mainly records and stores information manually, which is not only inefficient, but also has problems of large errors and high costs in the process of information statistics [20]. The introduction of digital carrier technology can improve the recognition rate of goods in the process of entry and exit, facilitate the visual management of goods in the warehouse, and also increase the transparency of the entire workflow in manufacturing logistics, improve the level of warehouse automation and operational efficiency of warehouse management, which is of great significance for further improving the level of enterprise logistics and warehouse management.

3.2. RFID based warehouse management system. RFID technology is a non-contact information transmission technology that identifies radio frequency signals through spatial coupling. It has long recognition

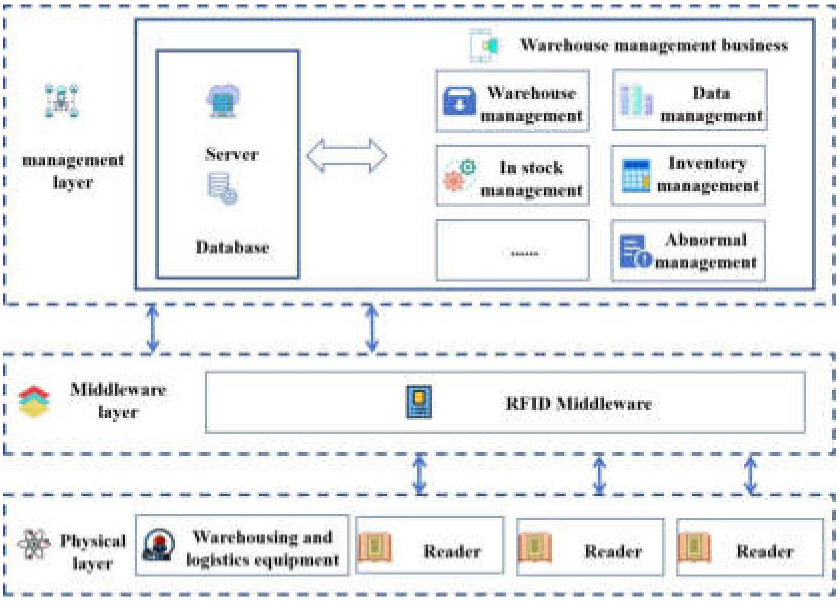


Fig. 3.1: RFID based warehouse management system

distance, small size, large data storage capacity, and can be reused. In the RFID system, the reader sends the signal from the tag through the transmitting antenna, and the carrier signal is demodulated and decoded before being sent to the receiving antenna. The data management system in the computer application software determines the validity of the tag through logical operations, and then processes and controls it according to different settings, issuing command signals.

The author designed an RFID based warehouse management system that includes management layer, middleware layer, and physical layer. The overall structural framework of the system is shown in Figure 3.1.

The management mainly controls the entire system process, is responsible for information management and equipment scheduling of the intelligent warehousing and warehousing management system, and completes functions such as querying, statistics, and scheduling of inbound operations.

The middleware layer is the integration of RFID data collection system and warehouse management system, responsible for storing the data collected by the physical layer in the database. The middleware layer transfers system information, receives scheduling instructions from the management layer, completes tasks based on the sent content, and timely transmits information to the physical layer, possessing strong implementation capabilities and processing efficiency.

The physical layer is responsible for deploying RFID devices and other logistics devices in inbound operations, mainly for data collection and rapid identification of cargo label information. It can receive instructions from the management and quickly transmit information from the work site.

3.3. Optimization of Intelligent Warehousing Warehousing. The warehousing process mainly involves taking inventory of the goods entering the warehouse, confirming specific data such as quantity and type of goods. The author takes the warehousing process of intelligent warehousing in the production and manufacturing workshop as an example to introduce the warehousing process.

3.3.1. Traditional warehousing process. Traditional warehousing and logistics management records goods at designated locations through manual recording. In this management mode, the space utilization of the warehouse cannot be planned in advance, and the accuracy and completeness of product information cannot be guaranteed through traditional records or barcodes. The time cost is too high, and the error rate in the process of recording information is high. The traditional warehousing process is shown in Figure 3.2.

Traditional warehousing methods can easily lead to location conflicts, increase material quantity errors,

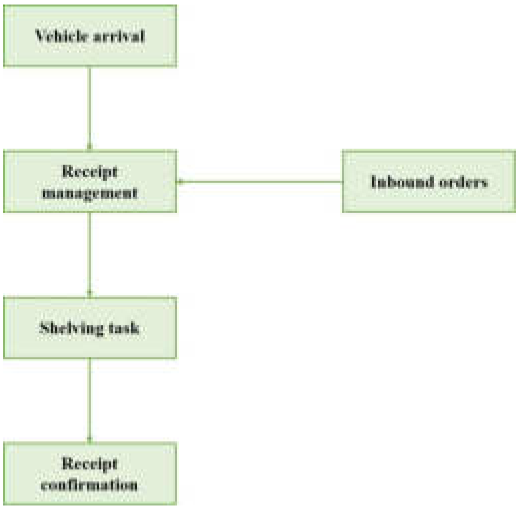


Fig. 3.2: Traditional Warehousing Process Diagram

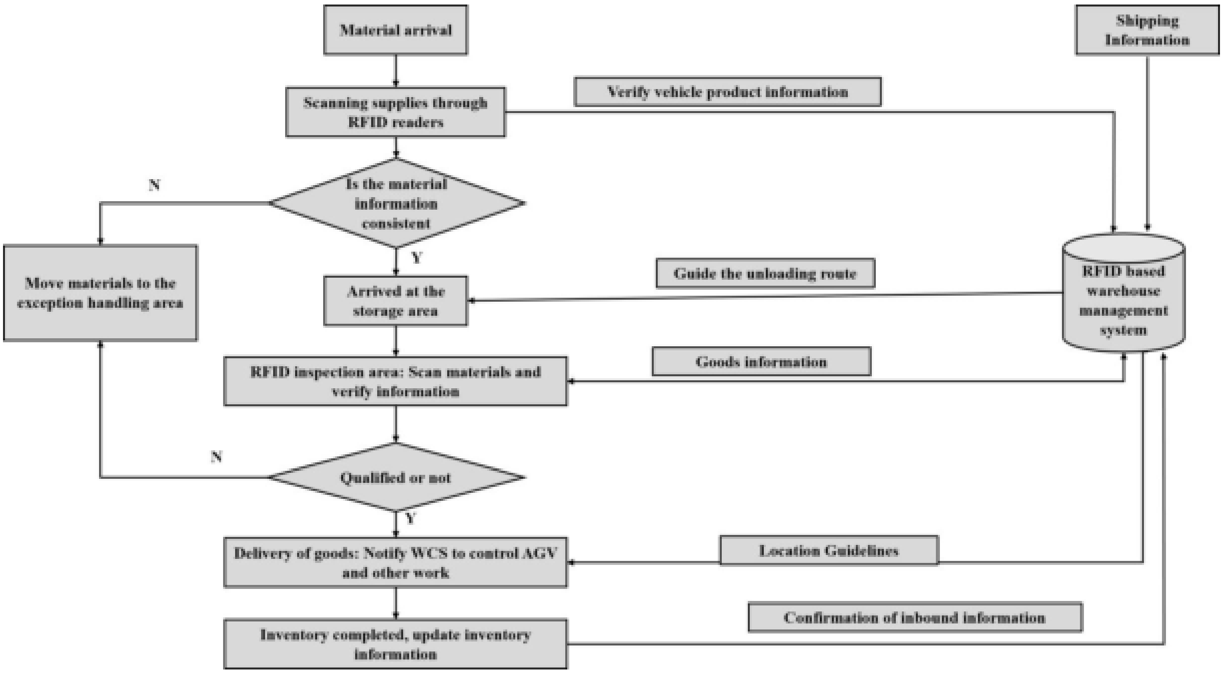

Material arrival
Shipping Information
Verify vehicle product information
Scanning supplies through RFID readers
N
Is the material information consistent
Y
Move materials to the exception handling area
Guide the unloading route
Arrived at the storage area
RFID based warehouse management system
Goods information
RFID inspection area: Scan materials and verify information
N
Qualified or not
Y
Delivery of goods: Notify WCS to control AGV and other work
Location Guidelines
Confirmation of inbound information
Inventory completed, update inventory information

Fig. 3.3: Inbound process diagram based on RFID warehouse management system

generate classification errors, and thus affect the operational efficiency of the entire warehousing management system.

3.3.2. Inbound process of RFID based warehouse management system. In response to the problems existing in the traditional warehousing process, the author has introduced digital carrier technology in logistics warehousing management to improve the warehousing process. The specific warehousing process is shown in Figure 3.3.

Enter vehicle information, cargo list, specific information, etc. into the computer system at the original warehouse of the goods, and generate relevant RFID tag information through computer software. Then, embed the tags into the packaging of each type of product. When the goods arrive at the storage area, they are scanned through an RFID reader, and the RFID warehouse management system synchronizes the information after receiving it. At the same time, WMS plans and allocates the storage area in the system. After the information of the goods is confirmed to be correct, they arrive at the storage area. The RFID warehouse management system scans the RFID tags on the items and generates target shelves based on the location allocation in the WMS system. After the goods arrive at the acceptance area, the reader at the entrance of the warehouse area scans and identifies the information of the incoming goods. The system will verify the data read with the cloud data, if the information is consistent with the database, send the relevant instructions to the WCS system. At this time, the destacking robot and AGV car move the goods to the designated location according to the system's path planning, and generate the inventory information in the WMS system.

If there is an abnormal reading of RFID information, the abnormal goods will be moved to the RFID exception processing area for processing. At the same time, the WMS terminal system in the exception processing area will provide the relevant information of the abnormal goods, including the warehouse receipt, to the exception handling personnel. The exception handling personnel handle the situation based on the abnormal information. After the exception handling is completed, the exception handling personnel can initiate a warehouse entry command on the WMS terminal system, and then the WMS system allocates a transfer robot to transport the goods into the warehouse.

3.4. Algorithm selection. The commonly used algorithms for logistics information data mining include neural network algorithms, genetic algorithms, fuzzy set algorithms, Bayesian algorithms, decision tree algorithms, and nearest neighbor algorithms. Below is a brief introduction and analysis of the Bayesian algorithm used in this article.

Bayesian algorithm is a general term for a class of classification algorithms based on Bayesian theorem, usually divided into naive, tree enhanced, and traditional Bayesian algorithms. Among them, Naive Bayes algorithm is the most common and easiest to implement among the three, and the author adopts this algorithm as the data mining algorithm for designing logistics information monitoring systems. The definition of this mining algorithm is as follows:

1. Assuming $A = \{a_1, a_2, \cdots, a_m\}$ is a raw dataset with m different feature attributes; $C = \{c_1, c_2, \cdots, c_n\}$ is a set of n different categories.
2. Use the set D of known classifications as the training sample set. Make category c and feature attribute a, and then calculate the conditional probability values of feature attribute a under category c.
3. Assuming that each feature attribute in A is conditionally independent, according to Bayesian theorem, it can be inferred that:

$$P(c_i|A) = \frac{P(A|c_i)P(c_i)}{P(A)} \tag{3.1}$$

Among them, molecules can be equivalent to:

$$P(A|c_i)P(c_i) = P(c_i)\prod_{j=1}^{m} P(a_j|c_i) \tag{3.2}$$

4. Calculate $P(c_1|A), P(c_2|A), \cdots, P(c_n|A)$ according to equation 3.1.
5. Find $P(c_k|A) = max\{P(c_1|A), P(c_2|A), \cdots, P(c_n|A)$, then $A \in c_k$.

The above process can be mainly divided into four stages: mining preparation, classifier training, classifier evaluation, and practical application. The main role of the mining preparation stage is to determine the feature attributes of the object to be mined and divide it manually, which has a significant impact on the processing effect of subsequent data; The training stage of the classifier is to use known training sample data to calculate the conditional probabilities of various feature attributes under different categories; The classifier evaluation stage calculates the category corresponding to the set A with the highest value for each category attribute, and obtains the corresponding classifier model; The practical application is to analyze the newly transmitted data based on the obtained model.

Table 3.1: Main Equipment Configuration Parameters

Name	Quantity	Main performance indicators
Storage area	2 storage ports	
Connecting device	2 connecting machines	Connect 8 boxes and 4 stacks at once
AGV	12	running speed $0.5m \cdot s^{-1}$
conveyor	Chain conveyor 2 locations	Conveying speed $1m \cdot s^{-1}$
Stacking robot	2	Complete the grabbing action in 4 seconds
Unstacking cache station	12	6 per warehouse entrance
Area of storage area	$1840m^2$	40m×46m

3.5. Simulation experiments. The author takes the logistics warehousing and storage area of the workshop as the design prototype, and uses the Flexsim simulation platform to fit and construct the data associated with each entity in the system in the order of the warehousing area, inspection area, and temporary storage area, based on the logistics distribution operation process sequence. A simulation model of the warehousing process is built. The main equipment configuration parameters of the model are shown in Table 3.1.

The design concept of the simulation model for the storage area is to divide the warehouse into three parts: upper, middle, and lower: The upper part includes two storage ports, two connecting devices, two cargo buffer areas, and an RFID exception handling area. The central area includes the temporary storage area for goods and the temporary storage area for empty shelves; The lower part includes two shelf storage areas. The AGV car spreads throughout the entire area according to the set position.

4. Results and Discussion. The workflow of the simulation system designed by the author based on RFID warehouse logistics management system is as follows: when raw materials arrive at the warehouse, the truck transports the goods to the entrance, and a double row chain conveyor is set up at the entrance to transport the goods. The dual track mobile connection device can connect 8 boxes and 4 stacks at once. The mobile connection device adjusts the height through photoelectric tubes and then connects with the chain machine of the material transport vehicle. The goods are placed on the conveyor and transported to the warehouse area. When the destacking robot is destacking, it grabs the goods and reads the RFID information at the end of the chain conveyor RFID reader. At the same time, the read information is transmitted to the RFID warehouse management system for vehicle and cargo information verification. After reading, the destacking robot places the goods on the destacking buffer station. When the system initiates the inbound command, the destacking robot grabs the goods from the cache station, reads RFID information, and binds it with the corresponding shelf barcode, transmitting the information to the inbound management system. The warehouse management system allocates storage locations and dispatches transfer robots to carry out material handling, achieving full automation in the process. If an RFID exception occurs during the process, the transfer robot will transport the box of goods to the RFID exception handling area.

Assuming that one unit of simulation time in the system is equivalent to the actual 1 second. In order to simulate the actual warehouse operation for 24 hours, set the simulation time to 86400 units. After setting the time, reset and run the model. After 24 hours of running, perform statistical analysis on the state data in the model. The simulation results are shown in Figures 4.1 to 4.3.

According to the simulation results, it can be seen that at the end of a simulation cycle, the storage volume at storage 1 and storage 2 is 2876 boxes and 2880 boxes, respectively. The work efficiency of the destacking robots at the storage port is 58.84% and 60.11%, respectively. The two destacking robots operate uniformly during storage operations and have a high utilization rate. From Figures 5 and 6, it can be seen that the total utilization rate of AGV in the simulation model is 72%, while the utilization rates of AGV11 and AGV12 are 36.70% and 18.64%, respectively, with utilization rates of less than 50%. This is because there are too many devices, resulting in low utilization of some devices. It can be inferred that idle equipment can easily lead to resource waste.

In practical system design applications, the number of AGVs running in the warehouse area can be set based on the order quantity of goods, reducing the idle running time of equipment, thereby improving the efficiency of the warehouse storage process, reducing energy consumption during idle running, and saving production costs.

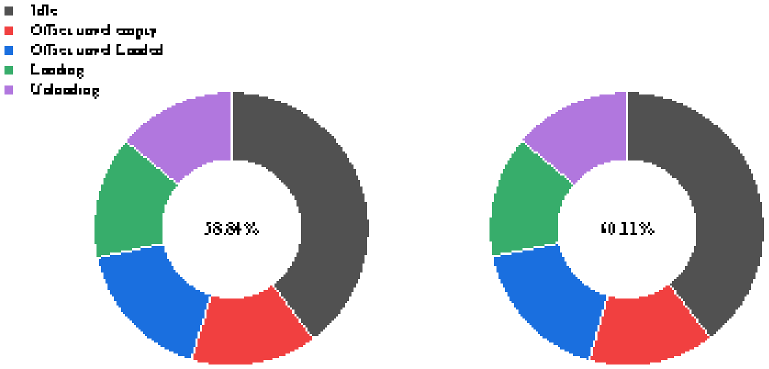


Fig. 4.1: Dynamic pie chart of utilization rate of inbound and destacking robots

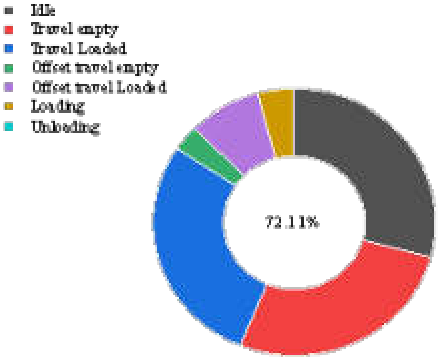


Fig. 4.2: Dynamic pie chart of total utilization rate of AGV in storage

The author used Flexsim's logistics operation simulation to virtually reproduce the actual process of inbound operations, which can optimize the model entity in a relatively short period of time. This result provides theoretical support for the operation and improvement of actual warehouses.

5. Conclusion. The author proposes the application of artificial intelligence in logistics management optimization research, focusing on the warehousing operations in production and manufacturing workshops, and applying RFID technology to warehousing management. The author constructs an RFID based warehouse management system architecture based on the design of management layer, middle layer, and physical layer, and applies it to the warehousing process of warehouse management, solving the problems of slow data collection and low accuracy in traditional logistics warehousing processes. The author used Flexsim software to establish a simulation model for warehousing, and through the establishment of various modules for warehousing, they intuitively understood the operation of the system, which has substantive guiding significance for the operation of a real warehousing and warehousing system.

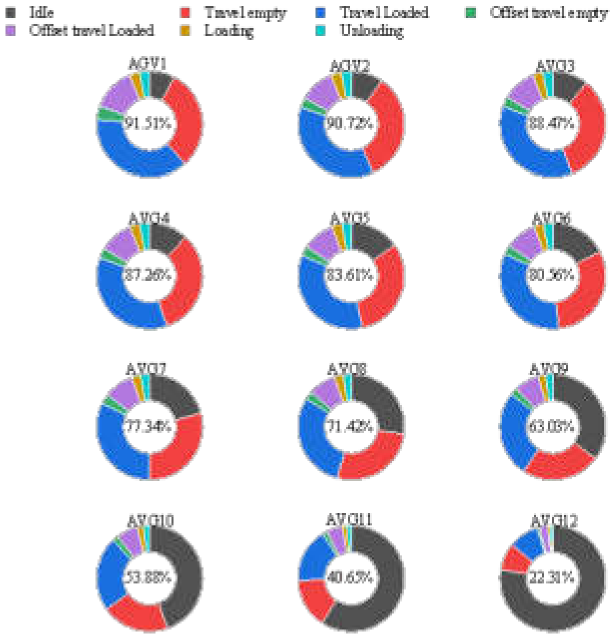


Fig. 4.3: Dynamic pie chart of utilization rates of various AGVs in storage

REFERENCES

[1] Li, R., & Yu, G. (2023). A complementarity model for a supply chain network equilibrium problem with electronic commerce. Journal of Industrial and Management Optimization, 19(9), 6705-6735.

[2] Ke, G. Y., & Bookbinder, J. H. (2023). Emergency logistics management for hazardous materials with demand uncertainty and link unavailability. Journal of Systems Science and Systems Engineering, 32(2), 31.

[3] Lehner, R., & Elbert, R. (2023). Cross-actor pallet exchange platform for collaboration in?circular supply chains. The International Journal of Logistics Management, 34(3), 772-799.

[4] Zhang, X., Yuan, J., Dan, B., Sui, R., & Li, W. (2022). The evolution mechanism of the multi-value chain network ecosystem supported by the third-party platform. Journal of Industrial and Management Optimization, 18(6), 4071-4091.

[5] Chen, W., Fan, J., Du, H., & Zhong, P. (2023). Investment strategy for renewable energy and electricity service quality under different power structures. Journal of Industrial and Management Optimization, 19(2), 1550-1572.

[6] Wei, J., Zhang, K., Cai, X., & Zhang, R. (2023). Study on the influencing factors of"internet+"recycling of waste mobile phone dominated by retailers, 11(3), 332-348.

[7] Wang, X., & Zhang, R. (2022). Carpool services for ride-sharing platforms: price and welfare implications. Naval Research Logistics NRL , 69(4), 550-565.

[8] Zhou, L., Fan, T., Zhang, L., & Chang, L. (2023). Quality differentiation with manufacturer encroachment: is?first mover always an advantage for retail platform?. Industrial Management & Data Systems, 123(3), 762-793.

[9] Yen, Y. S. (2023). Channel integration affects usage intention in food delivery platform services: the mediating effect of perceived value. Asia Pacific Journal of Marketing and Logistics, 35(1), 54-73.

[10] Parry, G., & Brookbanks, M. (2022). The impact of a blockchain platform on trust in established relationships: a case study of wine supply chains. Supply Chain Management: An International Journal, 27(7), 128-146.

[11] Liu, S., Hou, J., & Li, Y. (2022). Research on business model optimization of artificial intelligence garbage classification, 9(1), 20-25.

[12] Gaber, Y. H., El-Khodary, I. A., & Abdelsalam, H. M. (2023). A model review on joint optimization of part quality inspection planning, buffer allocation, and preventive maintenance in smms. Journal of Advanced Manufacturing Systems, 22(03), 667-691.

[13] Pasciolly, R. M. R. J., & Laksono, S. (2024). Optimization of arterial stents for may-thurner syndrome management in west java: experience and outcome. Research in Cardiovascular Medicine, 13(1), 1-5.

[14] Liu, C., Tang, C., & Li, C. (2023). Research on delivery problem based on two-stage multi-objective optimization for takeout riders. Journal of Industrial and Management Optimization, 19(11), 7881-7919.

[15] Zhan, J., Dong, S., & Hu, W. (2022). Ioe-supported smart logistics network communication with optimization and security.

Sustainable Energy Technologies and Assessments(Aug. Pt.A), 52.

[16] Yang, X. (2022). Optimization algorithm of logistics transportation cost of prefabricated building components for project management. Journal of Mathematics, 2(2), 244-263.

[17] Zhao, H., & Sharma, A. (2023). Logistics distribution route optimization based on improved particle swarm optimization. Informatica, 27(2), 303-314.

[18] Xu, X., Li, H., Xu, W., Liu, Z., Yao, L., & Dai, F. (2022). Artificial intelligence for edge service optimization in internet of vehicles: a survey. Tsinghua Science and Technology, 27(2), 270-287.

[19] Yuan, Z., Zhang, X., Li, H., Shen, P., Wen, J., & Wang, Z. L., et al. (2023). Enhanced performance of triboelectric mechanical motion sensor via continuous charge supplement and adaptive signal processing, 16(7), 10263-10271.

[20] Orjuela, K. D., Leppert, M. H., & Carroll, J. D. (2024). Navigating the gray: the complex story of pfo closure utilization. Circulation: Cardiovascular Quality and Outcomes, 17(1), 10581-27.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 5, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1276–1283, DOI 10.12694/scpe.v26i3.4295

LOGISTICS PATH PLANNING BASED ON IMPROVED PARTICLE SWARM OPTIMIZATION ALGORITHM

LONGJIAO TANG[*]

Abstract. In order to solve the problem of vehicle path scheduling management more reasonably, the author proposes a logistics path planning research based on improved particle swarm optimization algorithm. The author introduces a particle swarm optimization algorithm that incorporates a dynamic monkey jumping mechanism. Initially, dynamic population grouping is used to assign varying dynamic inertia weights, enhancing the algorithm's speed. Subsequently, the monkey jumping mechanism is added to ensure global convergence. This enhanced algorithm was then tested on two logistics distribution path optimization scenarios. In a consistent environment, the improved algorithm outperformed the standard particle swarm optimization algorithm by achieving a better optimal path fitness value, shorter average operation time, and a higher number of successful attempts to find the optimal solution. The experimental results show that out of 10 instances solved using the improved algorithm, 5 times obtained the optimal solution of 67.1km, and the optimal delivery path corresponding to the optimal solution was 0-4-7-6-0; 0-2-8-5-3-1-0, with an average calculation time of 1.26s, indicating high computational efficiency. The total delivery distance and average calculation time of the particle swarm algorithm, as well as the number of times to obtain the optimal solution, are 69.01, 2.7, and 3, respectively. It is evident that the enhanced particle swarm optimization algorithm significantly outperforms the conventional particle swarm algorithm. The improvements not only accelerate the optimization process but also enhance the algorithm's convergence, ensuring high-quality optimization results. Consequently, this improved algorithm holds substantial application value.

Key words: Particle swarm, Logistics distribution, Monkey jumping, weight coefficient

1. Introduction.

In today's rapidly growing era of Internet e-commerce, logistics services have increasingly become an essential part of daily life [1]. Delivery, as the most important link in the entire logistics service process, has become a key goal pursued by various logistics companies with high performance time and low cost loss. A reasonable logistics distribution path can not only strengthen the core competitiveness of enterprises, but also provide a catalyst for the better and faster development of society [2].

The development of the logistics industry has become an important factor affecting the growth of a country's GDP. Reasonable planning of logistics performance plans to reduce distribution costs will bring huge economic and social benefits to the country [3-4]. In contemporary logistics, distribution serves as a crucial link that directly connects to consumers and represents the highest cost component in the entire logistics service process. Efficient distribution path planning, characterized by ”high timeliness service and low cost loss,” directly enhances the core competitiveness of enterprises. In today's logistics management, an optimized distribution route can significantly improve user experience through faster fulfillment times, while also reducing operational costs for enterprises by effectively managing distribution expenses; Optimizing logistics distribution paths can also help address societal issues like emission pollution and road congestion. This optimization fosters the alignment of resources, environment, and value within the country, promoting comprehensive sustainable development. Consequently, logistics path planning optimization has increasingly become a major research focus, attracting significant investment from experts, scholars, consulting firms, and related enterprises [5].

2. Literature Review.

The problem of logistics vehicle path planning was first proposed in 1959 and has received great attention and extensive research from scholars and experts in various fields such as logistics science, operations research, combinatorial mathematics, and network planning [6]. In recent decades, experts across various fields have made substantial research advancements, exploring numerous problem scenarios and models. Consequently, a variety of solution methods have continuously emerged. Cai et al. introduced a heuristic elastic particle swarm optimization algorithm. In this approach, the A* algorithm offers global guidance for path planning in large-scale grids, while the elastic PSO algorithm employs contraction operations to

[*] ChongQing College Of Finance And Economics, Chongqing, 402160, China (Corresponding author, LongjiaoTang8@163.com)

identify the global optimal path from local optimal nodes, allowing particles to converge rapidly. In addition, during the iteration process, the rebound operation ensures the diversity of particles. The author maintains particle diversity through a rebound operation. Computer simulations and experimental results demonstrate that this algorithm not only overcomes the A* algorithm's limitation of not generating the shortest path, but also avoids the issue of failing to converge to the global optimal path due to a lack of heuristics [7]. Wang et al. proposed a dual layer search particle swarm optimization algorithm. This algorithm groups particle populations and variable dimensions, derives sub objective functions, constructs a double-layer search space, defines dynamic parameter matrices, and achieves information exchange and spatial reconstruction in the subspace. For optimization problems with complex features, double-layer search can achieve alternating fine search and rough search, thereby reducing computational costs and improving optimization efficiency [8]. Sun, H. et al. proposed a source optimization (SO) method that combines particle swarm optimization with a genetic algorithm (PSO-GA). This hybrid approach iteratively determines the optimal intensity distribution of the source. In this method, the pixelated source is decoded as an optimization variable for the optimal value function in the SO model. The PSO-GA algorithm, as an effective hybrid solution, converts the discrete SO problem into an optimal search for the value function, thereby enhancing lithography imaging performance in reverse [9].

The main goal of the author is to provide reliable support for logistics enterprises in route decision-making solutions, achieve cost reduction and efficiency improvement, and thereby enhance their market competitiveness. In recent years, many scholars have used particle swarm optimization to optimize problems such as neural network training and power system control, while the vehicle path planning problem, as a classic problem in the field of combinatorial optimization, has been relatively less optimized by this technology. Therefore, based on the establishment of a mathematical model for vehicle routing problems, the author proposes an improved particle algorithm that can effectively improve algorithm performance to solve the problem, ultimately improving solution quality and reducing logistics and distribution costs. The author's research not only enriches the solution schemes for optimizing logistics path planning problems, but also provides a foundation for future scholars to continue studying problems in this field, with rich theoretical and practical significance.

3. Method.

3.1. Introduction to Particle Swarm Optimization Algorithm.

Particle Swarm Optimization (PSO) is a stochastic intelligent optimization algorithm inspired by population behavior, first introduced by Eberhart and Kennedy. This algorithm, categorized as a natural evolutionary algorithm, was developed by mimicking the swarming behaviors of organisms like insects, birds, animals, and fish [10-11]. After research, it was found that these biological populations share the common characteristic of being able to search for food in a certain cooperative way, and continuously evolve and change their search patterns by learning their own experience and the experience of other individual members, ultimately moving closer to the region containing the global optimal solution. The particle swarm optimization algorithm has demonstrated its superiority in solving practical problems due to its simple steps and low implementation cost. It has gradually become a key focus and research object for scholars in the field of intelligent optimization.

3.2. Process of Particle Swarm Optimization.

Initialization: Start by randomly assigning speeds and positions to a population of particles while setting necessary parameters [12].

Evaluation: Assess the fitness value of each particle in the population, initiating the evolutionary process.

Local Optimization: Evaluate the fitness of each particle and compare it with its historical best fitness. If an improvement is found, update the particle's position accordingly.

Global Optimization: Compare each particle's current best fitness with the global best fitness. If a particle achieves a better fitness, update the global best position.

Velocity and Position Update: Adjust the velocity and position of each particle based on its previous state and the best positions found so far.

Termination Check: Check if the termination condition is met. If not, repeat steps 2 to 5. If yes, end the process. This process forms the basic flow of the PSO algorithm, aiding in understanding its operation (Figure 3.1).

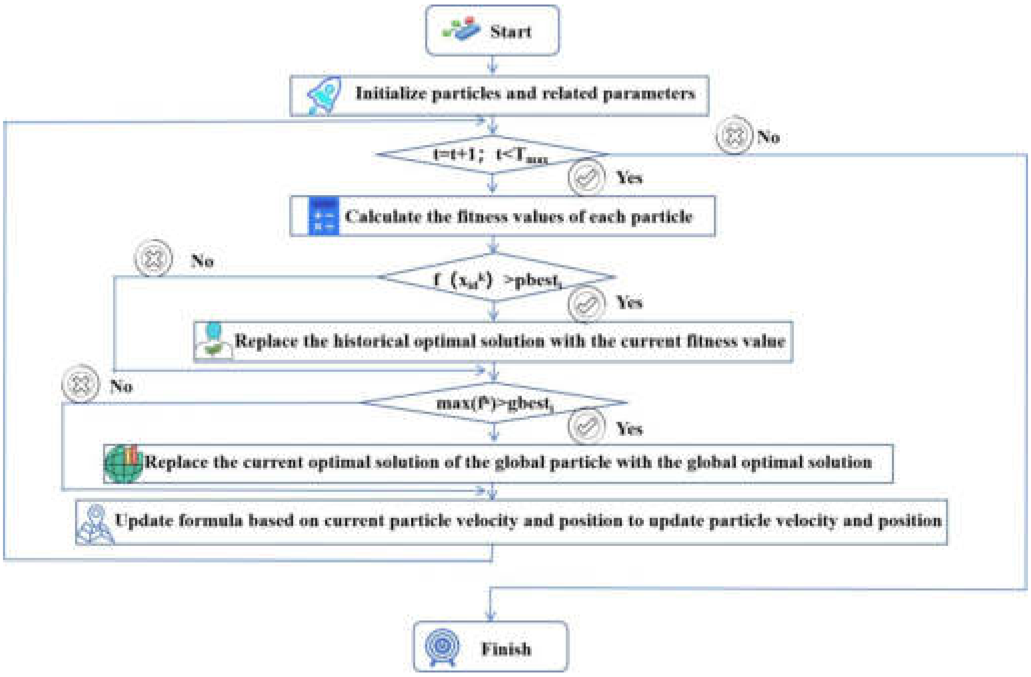

Start
Initialize particles and related parameters
t=t+1; t<T_max
No
Yes
Calculate the fitness values of each particle
No
f (x_id^k) >pbest_i
Yes
Replace the historical optimal solution with the current fitness value
No
max(f^k)>gbest_i
Yes
Replace the current optimal solution of the global particle with the global optimal solution
Update formula based on current particle velocity and position to update particle velocity and position
Finish

Fig. 3.1: Flowchart of the original PSO algorithm

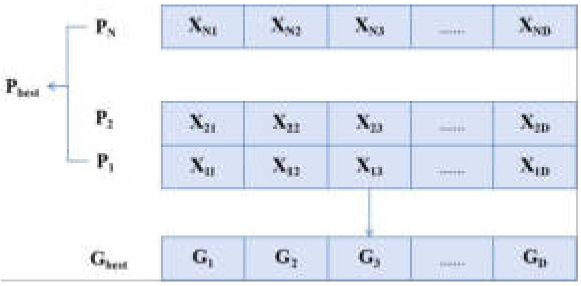


Fig. 3.2: Pbest Guided Gbest’s Dimensional Learning Strategy Model

3.3. Dimension wise learning strategies. At present, in most studies, the strategy of selecting all dimensions to update and re evaluate the optimal particle guidance in the population as a whole is adopted. However, for complex multi-dimensional function optimization problems, using this method may mask the information of certain correct evolutionary dimensions due to the interference between dimensions, resulting in a waste of evaluation times and reducing the convergence speed and efficiency of the algorithm [13]. The dimension by dimension learning strategy separates the optimal solution and the learning object in dimensions, independently examining the information in each dimension, which can effectively avoid the problem of inter dimensional interference. In PSO, as the population evolves, the Pbest of each particle is constantly updated, recording and updating its historical best performance during flight, with high utilization value. Therefore, in order to ensure the diversity and effectiveness of learning objects in the dimension wise strategy, the author proposes a Pbest guided Gbest dimension wise learning strategy based on the characteristics of Pbest and the advantages of dimension wise strategy.

Figure 3.2 shows a schematic diagram of the model for this strategy. The figure introduces the push and push operations in the data structure to simulate the actions of all Pbest in the population to guide the optimal particles one by one.

The idea of this strategy is to decompose the position vectors of Gbest and Pbest by dimension, and combine the values of one dimension on Pbest with the values of other dimensions on Gbest to form a new Gbest; New solution for evaluating fitness values; If the current new solution has better quality, retain the update result of Pbest dimension information on the solution; Otherwise, abandon the current dimension value and keep the original Gbest dimension information unchanged. Adopt this greedy evaluation method until all dimensions are updated. After all the dimensions of a Pbest have been guided by the corresponding dimensions of Gbest, a stack exit operation is performed to leave the Pbest stack container and compress it, starting the guidance of other Pbest on Gbest [14].

By introducing a greedy evaluation strategy into the dimension by dimension learning strategy, the degradation of certain dimensions is completely eliminated, avoiding the problem of masked evolutionary dimension information, and thus obtaining higher quality solutions, significantly improving convergence accuracy. Meanwhile, unlike most dimension wise learning strategies that use the method of learning from a single object, the Gbest in this strategy is influenced by the guidance of the population individual Pbest, strengthening the connection between the individual optimal particle and the population optimal particle, and improving the diversity of the learning objects for the optimal particle [15,16].

3.4. Corrective Strategies. In traditional PSO, particles are updated according to their velocity and displacement during the evolution process. They are guided by individual best and group best during the evolution process, lacking attention to the entire motion process of particles. Especially in complex multimodal functions with high optimization difficulty, particles generate a lot of randomness during the evolution process, which is one of the important reasons for the slow convergence speed of particle swarm optimization algorithms. During population evolution, particles might initially move towards the optimal solution direction. However, due to the intricacies of optimization, subsequent generations might deviate from this path. At this time, if we continue to follow the update speed and displacement, guided by the error information generated by the random flight of some particles from the previous generation, it will inevitably waste particle learning time, leading to a slower convergence speed. In order to address this issue, the author proposes a correction strategy that intervenes in the optimization direction of the next generation of particles by monitoring the changes in their motion direction throughout the evolution process, in order to avoid further erroneous guidance and improve the convergence speed of the population. Figure 3.3 provides a simple schematic diagram of the correction strategy. The A-class particles in the figure represent particles that have been affected by randomness and incorrect guidance. The next generation update will deviate from the direction of the optimal solution, and the update speed will be reversed by taking the direction of the velocity vector, so that the next generation can move towards the direction of the optimal solution and improve the convergence speed; Meanwhile, B-class particles with the correct direction of motion will continue to be updated in the original way [17].

In the algorithm, each particle represents a potential solution to the problem, and the optimal solution is derived by considering both the current and historical best solutions. During each iteration of the particle swarm, particles update themselves based on the best solutions they have encountered. At each iteration, the velocity of the particles is updated using equation 3.1:

$$v_{ik}^{g+1} = Wv_{ik}^{g} + \sigma_1 r_1 (S_{ak}^{g} - S_{ik}^{g}) + \sigma_2 r_2 (S_{bk}^{g} - S_{ik}^{g}) \tag{3.1}$$

In the formula: σ_1 and σ_2 are the initialization acceleration constants; g is the particle swarm update algebra; In addition, r_1 and r_2 are two mutually independent random functions; w is a weight factor variable; v_i is the running speed of the current particle. When the particle swarm is iterated each time, its position is updated by equation 3.2:

$$S_{ik}^{g+1} = S_{ik}^{g} + v_{ik}^{g+1} \tag{3.2}$$

3.5. Establishment of a mathematical model for optimizing logistics distribution paths. If we want to deliver agricultural products to customers, assuming that the distance between customer i and customer j is represented by d(i,j) as i,j=0,1,$\cdots$M, and d(0,0) as the distribution center, based on the description of the logistics distribution path selection problem mentioned above, a mathematical model can be established as

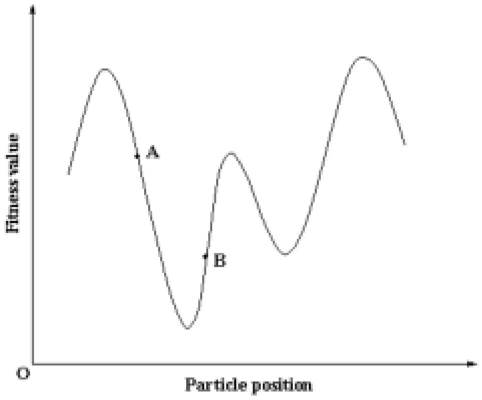


Fig. 3.3: 2D schematic diagram of correction strategy

follows:

$$Rou = \sum_{i=1}^{V}(\sum_{i=1}^{T} d(r_v^{j-1}, r_v^j) + d(r_v^T, 0)) \cdot sgnT \tag{3.3}$$

In the formula, r_v^j represents the order of customers in the vehicle v delivery path as j; When referring to the variables T and V, T represents the total number of customers transported by a specific vehicle (v), while V denotes the total number of vehicles. If T equals 0, it indicates that the vehicle has no customers to transport, whereas a value of 1 for T signifies that the vehicle has customers to transport. The parameter sgn (T) satisfies equation 3.4:

$$sgn(T) = \begin{cases} 1 & (T \geqslant 1) \\ 0 & (T = 0) \end{cases} \tag{3.4}$$

The constraint conditions for route optimization are:

$$\begin{cases} C_{v1} \cap C_{v2} = \phi, v_1 \neq v_2 \\ \sum_{i=1}^{T} q_{r_v} \leqslant Q_v, T \neq 0 \\ \bigcup_{v=1}^{V} R_v = \{1, 2, \cdots, M\}, 0 \leqslant T \leqslant M \\ \sum_{j=1}^{T} d(r_v^{j-1}, r_v^j) + d(r_v^T, 0) \leqslant L_v, T \neq 0 \end{cases} \tag{3.5}$$

Its optimization objective function:

$$G_{min} = \partial Rou^{\beta} \tag{3.6}$$

In the formula, C_v represents the set of customer points for vehicle v delivery; M is the total number of customers; q_i is the demand for customer i; Q_v is the maximum load capacity of vehicle v; R_v is a collection of customer points for vehicle v delivery; ∂ is the weighted coefficient; β is the amplification factor; L_v is the maximum distance transported by the vehicle; v_i is the set of vehicles that need to be used to transport customers [18]. According to equation 3.6, it can be concluded that logistics distribution not only requires fewer delivery vehicles, but also the shortest delivery path. It also requires delivering goods to customers within a specified time. This essentially involves finding an optimal logistics distribution route that satisfies multiple

Table 3.1: Environmental Data

Serial Number	longitude/(°E)	latitude/(°N)	elevation/
0	117.2090576	39.11549092	1.635
1	117.2091435	39.11549092	1.750
2	117.2092296	39.11549092	1.957
3	117.2093154	39.11549092	2.014

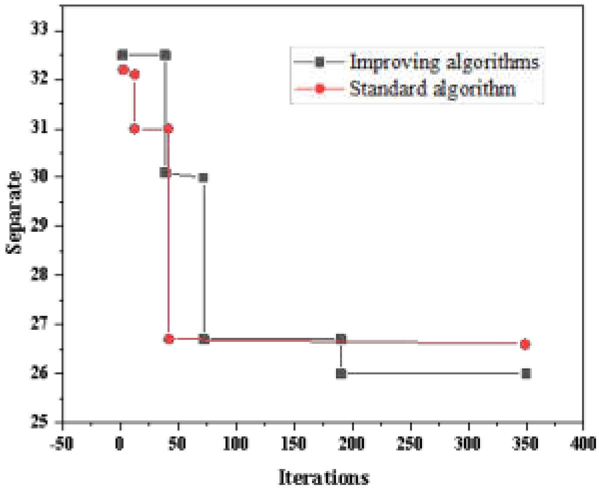


Fig. 4.1: Comparison of convergence speed between two algorithms

constraints simultaneously. The current particle swarm optimization algorithm simulates the foraging behavior of bird flocks and has parallel search capabilities. However, it has certain shortcomings, such as being trapped in local optima and slow maturation speed. In order to obtain a better logistics distribution path, it must be improved.

3.6. Simulation Environment and Parameter Settings. Use Google Earth to obtain geographic elevation data around a scenic area, and use ArcGIS to convert the features into grids as the logistics and transportation environment. Partial environmental elevation data is shown in Table 3.1.

Experiment 1. Logistics company has a distribution center with coordinates (0,0), which needs to deliver goods to customers at point coordinates (15,15). There are some buildings between them as shown in Table 1. The experimental hardware for simulation is: P5 dual core 2.9GCPU, 4GB memory, and 500G hard disk; The operating system is Windows 7 and the programming language is VC6.0++.

4. Results and Discussion. In the same environment, when comparing the convergence speed, the improved particle swarm algorithm's convergence process, as depicted in Figure 4, is noteworthy. Compared to the standard particle swarm algorithm, both algorithms exhibit rapid progress towards local optimal solutions. However, with increasing iterations, the standard particle swarm algorithm struggles to escape local optima. Conversely, the improved particle swarm algorithm, incorporating a dynamic monkey jumping mechanism, proves superior by generating 25 better solutions when the iteration count reaches 186. Hence, this mechanism effectively addresses the limitations of the standard particle swarm algorithm [19].

To evaluate the superiority of the enhanced particle swarm algorithm, we conducted comparative simulations with three other optimization algorithms: particle swarm algorithm, genetic algorithm, and ant colony

Table 4.1: Calculation Results of Improved Algorithm

Calculation times	1	2	3	4	5	6	7	8	9	10	average
Total delivery distance/km	68.3	68.3	67.1	67.7	68.3	68.1	67.1	67.1	67.1	68.2	67.82
Calculation time/s	1.1	1.0	1.0	1.4	1.3	1.7	1.3	1.0	1.0	1.2	1.26

algorithm. Each algorithm underwent 100 runs, and the average outcome was considered the final result. Here are the simulation test results indicating the optimization success rates for each algorithm:Particle Swarm Algorithm,Genetic Algorithm,Ant Colony Algorithm,Improved Particle Swarm Algorithm.These results provide insights into the relative performance of each algorithm, highlighting the effectiveness of the enhanced particle swarm algorithm: The number of failures is 12, 5, 7, 3, the success rates are 88%, 95%, 93%, 97%, and the running time is 10, 4.2, 5.2, and 1.8 seconds, respectively. The comparison results show that the improved particle swarm optimization algorithm has increased the success rate of logistics distribution path optimization, increased the speed of the algorithm, and effectively prevented other algorithms from falling into local optima and premature convergence defects.

Experiment 2. A certain delivery problem requires delivery from a central warehouse to 8 customers. There are 2 vehicles in the central warehouse, each with a load capacity of 8 tons. The maximum driving distance for each delivery is 40 km, and the demand of each distributor is set to qi (i=1,2... 8) (unit: t). The distance between the distribution center and each distributor, the demand of customers, and the distance between the distribution center and customers, as well as between customers and customers (0 represents the distribution center, 1-8 represents 8 customer points). The parameter settings are the same as above, and the improved algorithm is used to run and solve 10 times. The calculation results of the total delivery distance and calculation time are shown in Table 4.1.

Out of 10 instances solved using the improved algorithm, 5 times obtained the optimal solution of 67.1km, and the optimal delivery path corresponding to the optimal solution was 0-4-7-6-0; 0-2-8-5-3-1-0, with an average calculation time of 1.26s, indicating high computational efficiency. The total delivery distance and average calculation time of the particle swarm algorithm, as well as the number of times to obtain the optimal solution, are 69.01, 2.7, and 3, respectively. The enhanced particle swarm algorithm clearly outperforms the standard particle swarm algorithm. Sensitivity analysis of the new algorithm reveals that the performance of traditional particle swarm optimization algorithms is greatly influenced by the choice of weight coefficients. Optimal weight coefficients typically fall within the range of [0.8, 1.2]. If the size of the weight coefficients is not selected properly, it directly leads to the algorithm not finding the appropriate optimal solution; The enhanced algorithm addresses the necessity for larger weight coefficients during the initial optimization stages to expedite convergence. As optimization progresses and approaches the optimal solution, weight coefficients gradually decrease, facilitating more precise solution refinement. This adaptive approach to weight coefficients influences the algorithm's performance across various optimization ranges and periods. However, compared to traditional particle swarm optimization methods, the improved algorithm demonstrates reduced sensitivity to the specific size of weight coefficients, indicating enhanced robustness and effectiveness. Through experiments, it is known that the reason why the improved particle swarm optimization algorithm outperforms other algorithms in terms of search performance and optimization path planning ability is mainly because the inertia weight factor of the improved particle swarm algorithm adaptively changes with the dynamic grouping of the group. Furthermore, the integration of the monkey jumping mechanism addresses the issue of ”premature convergence,” enhancing the efficiency of logistics distribution path exploration. This improvement enables faster discovery of optimal logistics distribution paths, thereby facilitating safe and efficient distribution in agricultural production processes. Simulation results affirm that the enhanced particle swarm optimization algorithm serves as an effective tool for optimizing logistics distribution paths [20].

5. Conclusion. The author introduces a study on logistics path planning using an enhanced particle swarm optimization algorithm. Addressing the limitations of existing heuristic algorithms in logistics distribution path optimization, the study proposes an algorithm leveraging an improved dynamic monkey jumping mechanism. Simulation outcomes demonstrate that this particle swarm optimization algorithm, enriched with a dynamic monkey jumping mechanism, enhances the success rate of logistics distribution path optimization, yielding superior solution outcomes and efficiency. These findings offer valuable insights for the exploration of other heuristic algorithms and logistics distribution path optimization challenges.

REFERENCES

[1] Wahyuni-Td, I. S. , Abideen, A. Z. , Fernando, Y. , & Mergeresa, F. . (2023). Traceability technology, halal logistics brand and logistics performance: religious beliefs and beyond. Journal of Islamic Marketing, 14(4), 1007-1031.

[2] Su, J. , Shen, T. , & Jin, S. . (2022). Ecological efficiency evaluation and driving factor analysis of the coupling coordination of the logistics industry and manufacturing industry. Environmental science and pollution research international, 29(41), 62458-62474.

[3] Manukjan, N. , Fulton, D. , Ahmed, Z. , Blankesteijn, W. M. , & Sébastien Foulquier. (2024). Vascular endothelial growth factor: a double-edged sword in the development of white matter lesions. Neural Regeneration Research, 20(1), 191-192.

[4] Alnpak, S. , Isikli, E. , & Apak, S. . (2023). The propellants of the logistics performance index: an empirical panel investigation of the european region. International Journal of Logistics Research and Applications, 26(7), 894-916.

[5] Ke, G. Y. , & Bookbinder, J. H. . (2023). Emergency logistics management for hazardous materials with demand uncertainty and link unavailability. Journal of Systems Science and Systems Engineering, 32(2), 31.

[6] Gao, Y. , Xia, J. , & Ke, H. . (2022). A branch-and-cut algorithm for hub network design problems with profits. Naval Research Logistics NRL , 69(4), 622-639.

[7] Cai, L. . (2023). Decision-making of transportation vehicle routing based on particle swarm optimization algorithm in logistics distribution management. Cluster computing(6), 59(17), 334-340.

[8] Wang, Y. , Lei, Z. , & Wu, J. . (2023). Bilevel-search particle swarm optimization algorithm for solving lsgo problems. Journal of ambient intelligence and humanized computing(12), 35(5), 109-109.

[9] Sun, H. , Zhang, Q. , Jin, C. , Li, Y. , Tang, Y. , & Wang, J. , et al. (2023). Inverse lithography source optimization via particle swarm optimization and genetic combined algorithm. IEEE Photonics journal, 52(2), 176-189.

[10] Xiong, X. , Huang, R. , & He, C. . (2022). Research on intelligent path planning technology of logistics robots based on giraph architecture. International Journal of Computing Science and Mathematics, 16(3), 252.

[11] Jiang, W. , Liu, S. , & Li, S. . (2023). An extended cross-efficiency evaluation method based on information entropy with an application to the urban logistics industry. Journal of Modelling in Management, 18(2), 578-601.

[12] ZHANG Guang-sheng. (2022). Study on quality incentive mechanism of logistics service supply chain under information asymmetry. Journal of Highway and Transportation, Research and Development, 39(5), 157-165.

[13] Fan, J. P. , Zhang, H. , & Wu, M. Q. . (2022). Dynamic multi-attribute decision-making based on interval-valued picture fuzzy geometric heronian mean operators. IEEE Access, PP(99), 1-1.

[14] Geng, S. , Wang, L. , Li, D. , Jiang, B. , & Su, X. . (2022). Research on scheduling strategy for automated storage and retrieval system, 7(3), 15.

[15] Zhang Haixia. (2022). Study on coal logistics supply chain management and transportation cost control method. Coal Economic Research | Coal Econ Research, 42(7), 60-64.

[16] Trejos, C. A. R. , Meisel, J. D. , & Jaimes, W. A. . (2023). Humanitarian aid distribution logistics with accessibility constraints: a systematic literature review. Journal of Humanitarian Logistics and Supply Chain Management, 13(1), 26-41.

[17] Wu, J. , Ding, W. , Zhang, Y. , & Zhao, P. . (2022). On reliability improvement for coherent systems with a relevation. Naval Research Logistics NRL , 69(4), 654-666.

[18] Sallns, U. , & Bjrklund, M. . (2023). Green e-commerce distribution alternatives – a mission impossible?for retailers?. The International Journal of Logistics Management, 34(7), 50-74.

[19] Jing, L. , Qianchao, L. , & Hao, L. I. . (2023). Uav penetration mission path planning based on improved holonic particle swarm optimization, 34(1), 197-213.

[20] LI Kai,XIAO Xi,DONG Shan-heng,SONG Xu. (2023). Research on intelligent scheduling based on improved particle swarm optimization. Manufacturing Automation, 45(2), 214-216.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 20, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1284–1292, DOI 10.12694/scpe.v26i3.4207

ALEX/ELM NETWORK DETECTION BASED ON IMPROVED FIREFLY SWARM OPTIMIZATION ALGORITHM

XIAOYAN WANG*

Abstract. To address the issues of blind spots and low detection accuracy associated with using a single machine learning approach in network intrusion detection, the author suggests employing an Alex/ELM network detection system enhanced by an optimized firefly swarm algorithm. In the construction of base classifiers, the differences between samples of each base classifier are increased by sampling the sample set and selecting the feature set; By employing various learning algorithms to boost the diversity of the base classifiers on the sample set, the detection results are combined through a weighting mechanism. An enhanced firefly optimization algorithm is used to fine-tune the classification result weights of each base classifier. Experimental results demonstrate that, compared to other algorithms, this approach maintains a relatively high detection accuracy (with a minimum accuracy of 95.5%), showcasing the algorithm's stability and effectiveness even with imbalanced samples. In conclusion, the method proposed by the author significantly enhances detection accuracy, while reducing both the false alarm rate and the missed alarm rate.

Key words: Intrusion detection, machine learning, heterogeneous integration, firefly optimization, Alex/ELM

1. Introduction. At present, at a time when a new round of scientific and technological industrial revolution and the risk of epidemic are intertwined and superimposed, the global network security situation is still grim. The frequent occurrence of network security threats against key industries, new technologies and new scenarios forces countries to continue to deepen security measures for key infrastructure and strengthen security risk prevention for new technologies and new applications. Cybersecurity legislation and law enforcement work together to fully defend cyberspace security [1]. Benefiting from policy intensification and the release of security needs, the development of network security technology has ushered in opportunities. Looking ahead to the 14th Five Year Plan period, the digital economy will shift towards a new stage of deepening development. In response to the new situation and challenges of network security, the connotation of network security concepts, new technologies, and security technology industries will usher in key developments [2]. The global cybersecurity situation remains strict, and the evolution and upgrading of network attack methods are not optimistic about the current global cybersecurity situation. On one hand, key industries like energy, transportation, and telecommunications have been increasingly targeted by cyber attacks. In 2021, numerous organizations were affected, including the largest oil pipeline operator in the United States, public railway operators in the UK, and major telecommunications companies in New Zealand. These attacks have caused multiple interruptions to network services and have had a profound impact on social stability and people's production and life [3]. On the other hand, there is an increasing number of network threats targeting new technologies and scenarios. Take the Internet of Vehicles as an example: the security risks associated with infrastructure components such as application platforms, network support, and data computing power are highly complex and multifaceted. According to the 2021 Global Automotive Network Security Report by Upstream, IoT infrastructure has emerged as a new target for cyber attacks. The percentage of connected vehicles subjected to cyber attacks rose to 77.6% from 2020 to 2021 alone. As network attack methods continue to evolve and improve, the conflict between network attack and defense has become increasingly intense [4]. Regarding attack methods, exploiting vulnerabilities to conduct chain attacks has become more common. In terms of tactics, enhanced network defense capabilities have made attacks more challenging. Consequently, attackers are now employing various strategies to bypass security measures and successfully infiltrate networks. When it comes to targets, driven by potential gains, attackers are increasingly selecting their targets with greater precision. By adopting intelligent methods, attackers begin to collect information about attack targets and target "high-value" targets to carry

[*] Luohe Medical College, Luohe, Henan, 462002, China (Corresponding author, XiaoyanWang7@163.com)

out attacks [5].

2. Literature Review.

The process of network intrusion detection is actually the recognition of abnormal network behavior in the system. When there is a significant difference between the current behavior and normal behavior, an alarm message is issued. This process was early achieved through traditional machine learning algorithms. According to whether there are data labels involved in the training process, traditional machine learning algorithms can be divided into supervised learning and unsupervised learning. According to different operating mechanisms, supervised learning can be divided into generative methods and discriminative methods. The generation method starts from a statistical perspective and uses probability distribution to reflect the similarity between traffic data. Representative algorithms include naive Bayesian algorithm, Bayesian network, and hidden Markov model. Chang, W. Y. et al. introduced a hybrid metaheuristic algorithm that integrates dynamic multi-swarm particle swarm optimization with the firefly algorithm. This approach aims to achieve an optimal deployment solution that maximizes coverage and minimizes energy consumption using both static and mobile sensors. Moreover, the proposed algorithm incorporates a novel switching search mechanism between subgroups to prevent early convergence from becoming trapped in local optima. The simulation results show that compared with other PSO based deployment algorithms, this method can achieve better solutions in terms of coverage and energy consumption [6]. Gao, B. T. et al. developed a method for self-correcting the parameters of a disturbance rejection controller using an enhanced firefly swarm optimization algorithm. This algorithm incorporates local optimization operators based on sine and cosine functions, along with adaptive mutation strategies. The refined algorithm is then applied to tune the parameters of the disturbance rejection controller, enhancing the control system's anti-interference capabilities and ensuring parameter accuracy. The results indicate that optimizing disturbance rejection control with the improved firefly swarm optimization algorithm results in a quick response time, no overshoot, a stable tracking process, strong anti-interference capability, and superior optimization performance [7]. Zhou, X. et al. developed a multi-objective optimization model utilizing an enhanced firefly algorithm. This model uses the partial load rate of each chiller unit and the cooling rate of the freezer as optimization variables to determine the ISAC system's minimal energy consumption loss rate and operating cost. Experimental results demonstrate that, compared to strategies based on constant proportion, particle swarm optimization, and the standard firefly algorithm, the optimization strategy based on the improved firefly algorithm (IFA) achieves significantly greater energy savings and economic benefits [8].

The author employs the firefly optimization algorithm to optimize the decision output weights of each base classifier, identifying the optimal weighting scheme to enhance the detection model's accuracy. By using ensemble learning methods as the central detection algorithm, the author enhances the detection model's performance by refining the construction of the base classifiers and the method of result fusion.

3. Research Methods.

3.1. Heterogeneous Integration Algorithm of GSO.

To enhance the effectiveness of ensemble learning, the author employs heterogeneous ensemble learning techniques, integrating multiple learning models and increasing the diversity of training samples from the training dataset. These methods improve the generalization capability of ensemble learning, thereby ensuring high detection accuracy.

3.1.1. Heterogeneous basis classifier generation.

Ensemble learning involves employing a finite set of learning machines to tackle the same problem, where the final output for a given input example is determined by aggregating the outputs of these machines within the ensemble. A prerequisite for ensemble classifiers to outperform individual classifiers is their individual accuracy and diversity. Current machine learning methods with robust generalization capabilities often meet accuracy requirements without stringent parameter considerations. The crucial focus lies in enhancing the diversity among base classifiers. Ensemble learning can be categorized into two methods based on the similarities and differences in the classification algorithms of the base classifier: isomorphic and heterogeneous ensembles. Isomorphic ensembles utilize the same learning algorithm across all base classifiers, varying only in parameters and selected samples. On the other hand, heterogeneous integration employs different classifiers and learning algorithms, effectively ensuring diversity among base classifiers. Hence, heterogeneous integration is adopted for investigating intrusion detection, emphasizing the importance of diverse classification methods in enhancing detection accuracy [9].

The differences between base classifiers are not only affected by the self classification algorithm mentioned above, but also by the selection of the dataset when constructing the base classifier. Ensemble learning can be classified based on dataset selection methods into Pattern Level ensemble, which involves various resampling techniques, and Feature Level ensemble, which focuses on selecting different sample features. Pattern Level ensemble utilizes methods like repeated sampling or altering sample distributions to create diverse training sets for each base classifier, enhancing their variability. Commonly used Bagging and Boosting methods fall under Pattern Level integration. Feature Level integration, on the other hand, targets scenarios with numerous features by selecting subsets that capture distinct problem properties for each base classifier's training set. Given the high-dimensional sample space in network intrusion detection, employing Feature Level integration is practical. To enhance base classifier accuracy, a heterogeneous ensemble construction method that combines Pattern Level and Feature Level approaches is employed.

3.1.2. Firefly Algorithm Weight Optimization. After acquiring the outputs from multiple base classifiers, it's essential to fuse these results to derive the ultimate detection outcome. Since intrusion detection revolves around binary classification, the output results of each base classifier are assigned as +1 and -1, representing normal and detected intrusion, respectively. Assuming there are n base classifiers, the detection result of the i-th base classifier is y_i, and the fused weight is x_i, the final detection result y is as follows 3.1:

$$y = sgn(\sum_{i=1}^{n} x_i y_i) \tag{3.1}$$

To ensure the model achieves its highest detection accuracy by determining the optimal x, the objective function for optimization is expressed as follows 3.2:

$$f(x_i) = max(acc), 0 < x_i < 1 \tag{3.2}$$

Among them, *acc* represents the testing and detection accuracy of the entire integrated model under different weights x_i. In order to solve this problem, the author chose the Firefly Optimization Algorithm (GSO) to solve it [10]. GSO is a typical swarm intelligence optimization algorithm, originally proposed by Krishnanand in 2005. Its basic idea is to simulate the movement of individuals with low brightness towards individuals with high brightness in firefly swarm activities, in order to achieve optimization.

Compared to other swarm intelligence algorithms, it has the advantages of simplicity, fewer parameters, and easy implementation, making it suitable for determining the weights of decision layers.

The current position of the i-th firefly in GSO, which has a weight of $x_i(t)$, has a fluorescence value of $I_i(t)$ at that position, and t is the number of iterations. The iterative update process of GSO is determined by both fluorescence and position updates.

The formula for updating the fluorescence value is as follows 3.3:

$$I_i(t) = (1-\rho)l_i(t-1) + \gamma f(x_i(t)) \tag{3.3}$$

Among them, $f(x_i(t))$ objective function fitness value, ρ is the volatilization factor of fluorescein, γ is the fluorescence renewal rate. The position update formula is as follows 3.4:

$$x_i(t+1) = x_i(t) + s(\frac{x_j(t) - x_i(t)}{||x_j(t) - x_i(t)}||) \tag{3.4}$$

Among them, s is the movement step size, $x_j(t) - x_i(t)$ is the distance between firefly j and i.

During each iteration, the dynamic decision domain radius is updated as follows 3.5:

$$r_d^i(t+1) = min\{r_S, max\{0, r_d^i(t) + \beta(n_i - |N_i(t)|)\}\} \tag{3.5}$$

Among them, r_S is the perception radius, β is the update rate, $|N_i(t)|$ is the number of fireflies within the neighborhood range. Through iteration, the optimal weight with the highest brightness will be ultimately found [11].

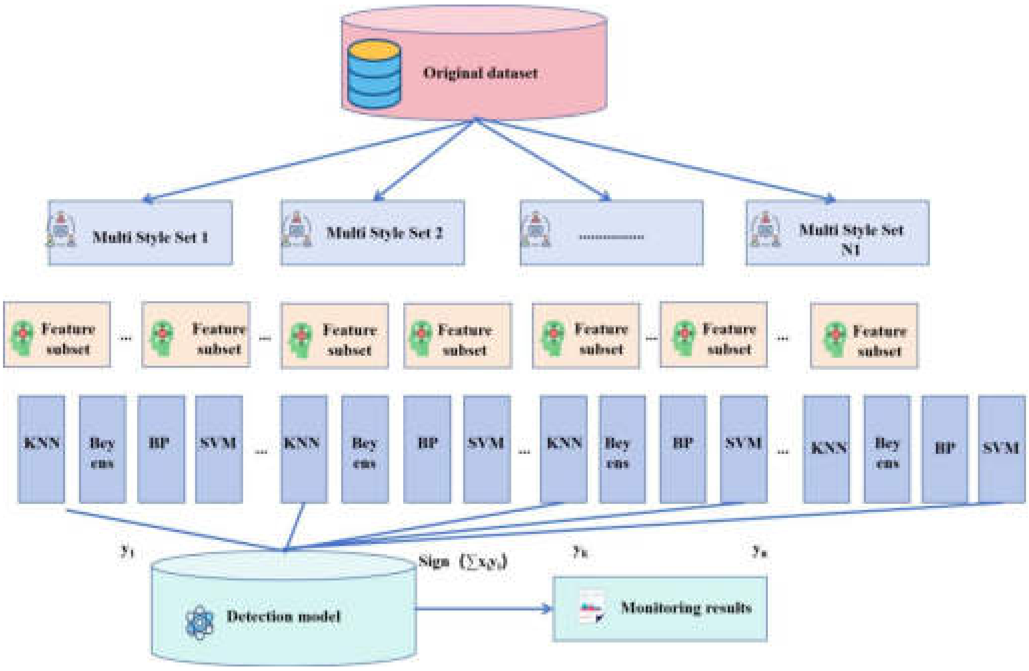


Fig. 3.1: Algorithm Process

3.1.3. Algorithm process. Building upon the previous analysis and considering the practicalities of intrusion detection problems, the author opts for heterogeneous integration. This involves selecting well-established machine learning methods with strong generalization performance: SVM, KNN, AlexNet, and ELM serve as learning techniques for base classifiers. To further improve the variability between base classifiers, an ensemble learning detection algorithm for GSO optimization weights combining Pattern-Level and Feature-Level selection training set methods. The detailed algorithmic process is depicted in Figure 3.1.

Here are the specific steps:

1. Employ the Bagging method to resample the samples with replacement, generating multiple subsets of samples.
2. Utilize the Feature Level method to randomly select features from the previously generated sample subsets, obtaining feature subsets.
3. Train different learning methods on the feature subsets obtained in Step 2. The parameters for each base classifier are determined using a simple trial and error method, resulting in the creation of each base classifier.
4. Synthesize the results from each base classifier with weighted fusion. The weights are optimized using the firefly algorithm as described in Section 1.2. The final detection results are then produced through a sign function [12].

3.2. Algorithm validation. In order to verify the effectiveness of the algorithm, we will first start with its effectiveness and examine its classification performance by applying the algorithm to the current universal dataset. Here, German, Ionosphere, Image, and Thyoid datasets are selected as the experimental datasets. The author's proposed algorithm is compared against traditional classification methods to evaluate performance differences. Additionally, a comparison is made between Bagging and Boosting ensemble algorithms to assess their respective effectiveness. For ensemble learning, it is not advisable to have too many or too few integrated base classifiers. Select 40 base classifiers and construct them based on this method, bagging, and boosting, respectively. The author samples the samples in the first layer and performs 5 resamples on each sample. Based on this, two random feature selections are performed on each subset, with a feature selection ratio of

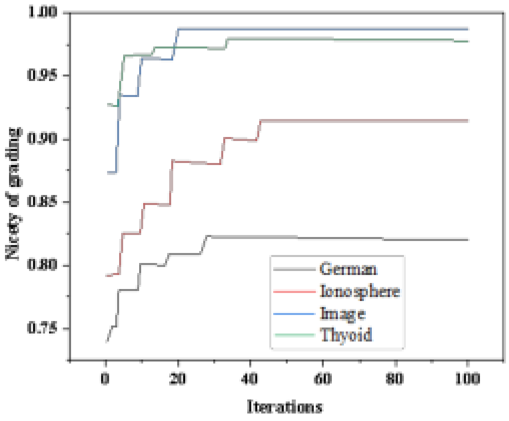


Fig. 3.2: GSO optimization results for different datasets

Table 3.1: Comparison of experimental accuracy for general datasets

data	Classification accuracy(Acc %)				
set	KNN	AlexNet	ELM	SVM	The method (Alex/ELM)
German	74.5	76.7	75.8	77.5	81.4
Ionosphere	81.3	82.7	84.5	86.2	90.7
Image	92.5	94.3	95.2	95.8	98.2
Thyoid	93.1	94.7	94.2	95.0	97.1

80%. This resulted in 10 training sets. Based on these 10 training sets, SVM KNN, AlexNet, and BP neural networks were trained to obtain 40 base classifiers. The final classification result is obtained using the majority voting method [13,14].

Firstly, compare the classification performance of a single classifier and this method. For the single classifier, we employ base classifier methods, namely SVM, KNN, AlexNet, and BP neural networks. In SVM and BP neural networks, Gaussian functions are chosen as the kernel and activation functions, respectively. The parameters for each learner are determined through 5-fold cross-validation and grid search methods. The parameters for each base classifier in this algorithm are straightforwardly set without any special processing. Given the robust learning capabilities of the selected methods, the chosen parameters typically fulfill the requirements. In the GSO optimization algorithm, based on the problem, set the number of fireflies to 50, the fluorescence volatilization factor =0.4, the fluorescence update rate =0.6, and the update rate =0.08, neighborhood threshold =5, with 100 generations. The obtained experimental results are shown in Figure 3.2.

The graph depicts that the classification accuracy of each dataset remains consistently high during the initial stages, suggesting that the ensemble learning algorithm yields promising learning outcomes and maintains stability throughout the process. On different datasets, the algorithms converge quickly, indicating that the GSO algorithm has good convergence, and each algorithm ultimately stabilizes at a relatively high classification accuracy [15]. Next, the results of this article will be compared with several other single classification methods, and Table 1 presents the classification results.

The experimental findings reveal that among the algorithms tested, KNN exhibits the lowest classification accuracy. Despite its simplicity and high operational efficiency, the KNN algorithm tends to extract minimal classification information. In contrast, SVM, AlexNet, and ELM yield comparable classification results. These methods leverage distinct learning mechanisms, offering varied perspectives for sample learning and demonstrating proficiency in handling nonlinear classification tasks. The author's proposed method achieves optimal

Table 3.2: Comparison of Average Differences of Base Classifiers

Data	Average difference value		
set	Bagging	Boosting	This method (Alex/ELM)
German	0.203	0.216	0.272
Ionosphere	0.170	0.181	0.2124
Image	0.013	0.012	0.020
Thyoid	0.064	0.073	0.103

Table 3.3: Classification accuracy results of integrated algorithms

Data	Classification accuracy(Acc%)			
set	Bagging	Boosting	Voting fusion	GSO optimization
German	78.4	78.4	80.3	81.3
Ionosphere	87.0	88.3	90.3	90.7
Image	97.3	97.1	97.7	98.1
Thyoid	95.3	95.8	96.4	97.1

classification accuracy by integrating diverse classifiers, thereby compensating for individual classifiers' errors and omissions. This ensemble learning strategy enhances the overall classification's generalization capability, showcasing the inherent advantage of ensemble learning.

Next, we'll assess the performance of this method against traditional ensemble learning algorithms. Traditional ensemble algorithms such as classic bagging and boosting integrate 40 base classifiers. The distinction among base classifiers serves as an indicator of ensemble learning effectiveness. The difference between base classifiers, denoted as D, is calculated using the following formula 3.6 on N samples:

$$Div = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} d_t(x_i) \tag{3.6}$$

Among them, the following equation 3.7:

$$d_t(x_i) = \begin{cases} 0 & if h_t(x_i) = f(x_i) \\ 1 & if h_t(x_i) \neq f(x_i) \end{cases} \tag{3.7}$$

Among them, $h_t(x_i)$ represents the predicted label of the t-th individual classifier on sample x_i, $f(x_i)$ is the predicted result after integrating all individual classifiers, using a difference threshold of at least 80% of the difference of the previous t individual SVMs, we evaluate the effectiveness of ensemble integration. A higher difference implies lower correlation between base classifiers, leading to improved integration outcomes [16]. We computed the average difference among three ensemble algorithms across various datasets, as summarized in Table 3.2.

Table 3.2 reveals that, on the whole, the disparity between Bagging and Boosting integration methods is not notably substantial. However, compared to the preceding methods, the average difference observed in this method is notably higher. This is because this method adopts heterogeneous integration. Below, we will examine the generalization performance of several ensemble algorithms by comparing their classification accuracy. Here, we'll delve into the GSO optimization employed by the author and the fusion decision results solely based on voting. Table 3.3 provides a comparison of the classification accuracy among these four methods.

From the experimental results, it can be seen that several ensemble algorithms have achieved certain improvements in classification accuracy compared to single learning methods. In the context of constructing the base classifier in this study, the classification accuracy of the voting fusion method surpasses that of traditional Bagging and Boosting. This improvement can be attributed to the increased diversity among base classifiers, leading to enhanced integration effects. Additionally, the GSO optimization method achieves the

Table 4.1: Results of Intrusion Detection Experiment

method	Normal detection number	noise factor	Attack detection count	false negative	accuracy
SVM	943	5.50	231	7.10	94.31
Bagging	973	2.50	236	5.10	97.31
Boosting	976	2.20	233	6.30	97.61
The method (Alex/ELM)	991	0.70	242	2.70	99.11

highest classification accuracy compared to simple voting methods [17,18]. These findings suggest that the method proposed in this article significantly enhances ensemble learning performance, thereby establishing a solid foundation for its application in intrusion detection.

4. Result analysis. The experiments conducted above validate the efficacy of the GSO-optimized ensemble learning algorithm proposed by the author in achieving high accuracy for binary classification tasks. This section extends the application of the algorithm to intrusion detection using the CSE-CIC-IDS2018 dataset. This dataset, collaboratively released by the Canadian Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC), comprises network traffic and system logs from 50 attack hosts targeting 420 computers and 30 servers across 5 departments within an enterprise. It encompasses a total of 7 types of attacks, including Brute force, Heartbleed, Botnet, DoS, DDoS, Web, and Infiltration. Compared to the traditional KDD99 dataset, CSE-CIC-IDS2018 can better simulate the network environment of enterprises, which includes a variety of network protocols and new attack methods. The original dataset is very large, with a ratio of approximately 4:1 between normal and invasive samples. In this experiment, training samples were selected in a 4:1 ratio. Specifically, 4000 normal samples and 1000 intrusion samples (including all 7 selected attack types) were randomly chosen from the training set. The test sample consists of 1000 normal samples and 250 attack samples. Individual machine learning methods, with SVM selected as the model, as well as ensemble learning algorithms such as bagging and boosting, are employed for detection experiments. It's worth noting that the base classifiers integrated with bagging and boosting algorithms are all SVM models. The performance metrics evaluated include the false detection rate, false alarm rate, and overall detection accuracy of each algorithm, as illustrated in Table 4.1.

The detection results presented in Table 4.1 highlight the comparatively high false detection and false alarm rates when solely employing the SVM algorithm for intrusion detection, underscoring the benefits of ensemble learning approaches. Among the three ensemble algorithms examined, the author's proposed algorithm demonstrates the most robust detection performance, aligning with our findings from experiments conducted on the UCI dataset. By enhancing the disparity between base classifiers, the algorithm's generalization performance is notably improved. Specifically, by analyzing the detection performance of different attack methods, we can observe that the algorithm proposed by the author is effective for Botnet The detection rate of attack types such as DoS, DDoS, and Web has basically reached 100%, and the main erroneous judgments are concentrated in the Brute force and Infiltration attack types. Analyzing the data of these two attacks, it was found that there are some normal traffic data in the abnormal samples, which have the same numerical values as the abnormal traffic in terms of characteristics. Comparing the performance of several methods on these two easily misclassified datasets, the method proposed by the author can greatly improve the detection accuracy of detection algorithms for these two types of attacks [19].

Moving forward, we'll utilize the Receiver Operating Characteristic (ROC) curve to evaluate the performance of the detection algorithms. When assessing a model's quality using the ROC curve, two key aspects are considered: the curve's shape and the Area Under Curve (AUC). A curve that approaches the upper left corner indicates superior detection performance, while a deviation from this corner suggests poorer performance. AUC represents the area beneath the ROC curve, serving as a reflection of the detection model's diagnostic value. A higher AUC value, closer to 1, signifies better model performance. Overall, the ROC curves of the algorithms employed earlier tend to cluster towards the upper left corner, indicative of their efficacy as detectors. Notably,

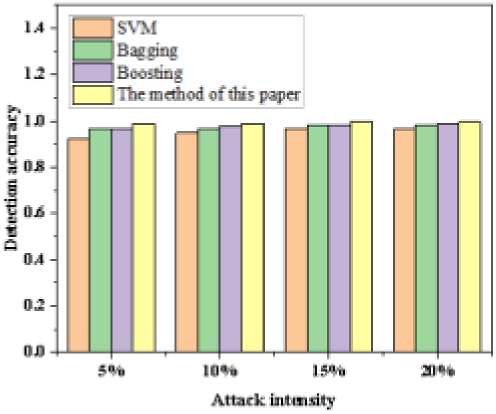


Fig. 4.1: Detection accuracy under different attack intensities

the author's proposed algorithm exhibits the most favorable bias, significantly outperforming other detection methods. Furthermore, in terms of AUC, this method boasts a larger area under the curve, reaching 0.992, surpassing the performance of the other three methods. In summary, the ROC curve analysis underscores the advantages of the method proposed by the author.

In real-world scenarios, network attacks often occur intermittently, with fluctuations in attack frequency over time. This variability is reflected in the dataset as the ratio of normal samples to attack samples. Consequently, collecting training datasets poses challenges, as a larger dataset theoretically leads to better detection performance. However, during the initial stages of detection, when a relatively small number of attack events are collected, the effectiveness of the detection model may be compromised. To assess the efficacy of detection models under different attack intensities, simulated scenarios are created with attack samples ranging from 5% to 20% of the total samples. The detection models are trained using these varying attack samples, and their detection accuracy under different learning modes is depicted in Figure 4.1.

Figure 4.1 illustrates that the trained detection models successfully identify intrusion events across varying attack intensities. When the attack intensity is low (i.e., a small proportion of attack samples in the training set, such as 5%), the SVM model achieves a detection accuracy of 91.4%. However, it exhibits a relatively high missed detection rate, primarily attributed to sample imbalance, causing the SVM's classification plane to deviate. From the experimental results, it can also be seen that sample imbalance can affect the effectiveness of machine learning. Compared with other algorithms, this method can consistently maintain a relatively high level of detection accuracy (with a minimum of 95.5%), indicating that the algorithm is stable and can also achieve good detection results in cases of imbalanced samples [20].

5. Conclusion. Given that network intrusion detection is essentially a binary classification problem with individual machine learning models often yielding suboptimal accuracy, a heterogeneous ensemble learning approach is adopted. By enhancing the disparity between the training sets and methods in the base classifiers, the overall integration effect is improved. Furthermore, to further enhance detection accuracy, enhancements have been made to the result fusion aspect of ensemble learning. The GSO algorithm is optimized to determine the optimal weight of the base classifier, and the final detection result is obtained through weighted fusion. Experimental results demonstrate the stability and accuracy of the proposed method, showcasing its practical value in real-world network intrusion detection applications. Addressing challenges associated with uneven training samples remains a focal point for future research efforts.

6. Acknowledgements.

1. The study was supported by Key scientific research projects of colleges and universities in Henan

Province“Research on the cultivation of College Teachers’ information teaching ability from the perspective of TPACK”(Grant No.22A880031);

2. The study was supported by Academic Degrees & Graduate Education Reform Project of Henan Province”Research on the path of graduate students’ Academic Integrity Construction”(Grant No. 2021SJGLX058Y).

REFERENCES

[1] Devaraj, A. F. S., Elhoseny, M., Dhanasekaran, S., Lydia, E. L., & Shankar, K. (2020). Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments. Journal of Parallel and Distributed Computing, 142, 36-45.

[2] Pitchaimanickam, B., & Murugaboopathi, G. (2020). A hybrid firefly algorithm with particle swarm optimization for energy efficient optimal cluster head selection in wireless sensor networks. Neural Computing and Applications, 32, 7709-7723.

[3] Igiri, C. P., Singh, Y., & Poonia, R. C. (2020). A review study of modified swarm intelligence: particle swarm optimization, firefly, bat and gray wolf optimizer algorithms. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 13(1), 5-12.

[4] Rahkar Farshi, T., & K. Ardabili, A. (2021). A hybrid firefly and particle swarm optimization algorithm applied to multilevel image thresholding. Multimedia Systems, 27(1), 125-142.

[5] Kaya, S., Gümüşçü, A., Aydilek, I. B., Karacizmeli, I. H., & Tenekeci, M. E. (2021). Solution for flow shop scheduling problems using chaotic hybrid firefly and particle swarm optimization algorithm with improved local search. Soft Computing, 25(10), 7143-7154.

[6] Chang, W. Y., Soma, P., Chen, H., Chang, H., & Tsai, C. W. (2023). A hybrid firefly with dynamic multi-swarm particle swarm optimization for wsn deployment. Journal of Internet Technology(4), 24-31.

[7] Gao, B. T., Shen, W., Dai, Y., & Ye, Y. (2022). Parameter tuning of auto disturbance rejection controller based on improved glowworm swarm optimization algorithm. Assembly Automation, 34(8), 6432-6440.

[8] Zhou, X., Yu, J., Zhang, W., Zhao, A., & Zhou, M. (2022). A multi-objective optimization operation strategy for ice-storage air-conditioning system based on improved firefly algorithm:. Building Services Engineering Research & Technology, 43(2), 161-178.

[9] Ab Talib, M. H., Mat Darus, I. Z., Mohd Samin, P., Mohd Yatim, H., Ardani, M. I., Shaharuddin, N. M. R., & Hadi, M. S. (2021). Vibration control of semi-active suspension system using PID controller with advanced firefly algorithm and particle swarm optimization. Journal of ambient intelligence and humanized computing, 12, 1119-1137.

[10] Kaya, S. (2023). A hybrid firefly and particle swarm optimization algorithm with local search for the problem of municipal solid waste collection: a real-life example. Neural Computing and Applications, 35(9), 7107-7124.

[11] Chou, D., & Jiang, M. (2021). A survey on data-driven network intrusion detection. ACM Computing Surveys (CSUR), 54(9), 1-36.

[12] Drewek-Ossowicka, A., Pietrołaj, M., & Rumiński, J. (2021). A survey of neural networks usage for intrusion detection systems. Journal of Ambient Intelligence and Humanized Computing, 12(1), 497-514.

[13] Ashiku, L., & Dagli, C. (2021). Network intrusion detection system using deep learning. Procedia Computer Science, 185, 239-247.

[14] Asif, M., Abbas, S., Khan, M. A., Fatima, A., Khan, M. A., & Lee, S. W. (2022). MapReduce based intelligent model for intrusion detection using machine learning technique. Journal of King Saud University-Computer and Information Sciences, 34(10), 9723-9731.

[15] Smys, S., Basar, A., & Wang, H. (2020). Hybrid intrusion detection system for internet of things (IoT). Journal of ISMAC, 2(04), 190-199.

[16] Mishra, N., & Pandya, S. (2021). Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. IEEE Access, 9, 59353-59377.

[17] Thakkar, A., & Lohiya, R. (2023). Fusion of statistical importance for feature selection in Deep Neural Network-based Intrusion Detection System. Information Fusion, 90, 353-363.

[18] Gümüşbaş, D., Yıldırım, T., Genovese, A., & Scotti, F. (2020). A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. IEEE Systems Journal, 15(2), 1717-1731.

[19] He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. IEEE Communications Surveys & Tutorials, 25(1), 538-566.

[20] Kumar, V., & Kumar, D. (2021). A systematic review on firefly algorithm: past, present, and future. Archives of Computational Methods in Engineering, 28, 3269-3291.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1293–1299, DOI 10.12694/scpe.v26i3.4209

ELECTRONIC INFORMATION IMAGE PROCESSING BASED ON CONVOLUTIONAL NEURAL NETWORKS

HONGMING PAN*

Abstract. In order to improve the accuracy and efficiency of mechanical part recognition, the author proposes a research on electronic information image processing based on convolutional neural networks. The author first performs saturation based grayscale processing on the image; Then, the binary image is obtained through significance enhancement, binarization using the Maximum Between Class Variance (OTSU) method, and morphological closure operation; Extract the part area using an improved seed filling method; Finally, the parts are identified by combining Scale Invariant Feature Transform (SIFT) features of image keypoints with Convolutional Neural Network (CNN) models. The experimental results show that the accuracy of the part recognition algorithm can reach 98.84%, and the recognition speed is about 5fps. Through experimental comparison and analysis, it has been proven that this method is fast and effective, with high accuracy and good robustness.

Key words: Part recognition, Image saturation, Seed filling method, Scale invariant feature transformation, Convolutional neural network

1. Introduction. With the popularization and development of science and technology, mobile phones, tablets, and computers have become necessities for people, whether it is work or life [1]. The emergence, development, and enrichment of technology have brought a new world to human civilization, in which electronic information provides convenient conditions for people to communicate in time and space, especially image processing technology, which is very widespread in daily life. At present, smartphones are extremely common, and users have a great demand for images. From the formation, acquisition, transmission, and reception of images, every step is inseparable [2-3]. However, in each stage, the image is more or less contaminated by noise, resulting in users not being able to achieve the expected image effect. However, directly optimizing or removing noise can also affect the accuracy of the image, so advanced noise removal techniques play a crucial role in the efficient use of images [4].

The use and transmission of image information is very common in today's era of electronic products, but the image information required by users is highly susceptible to external signal interference, so image denoising technology needs to be continuously improved [5]. The author described image denoising technology and superpixel generation algorithm, and explained the definitions, principles, and operational processes of these two aspects from a professional perspective, in addition, the author also analyzed the advantages and disadvantages of image denoising and superpixel generation algorithms, continued to innovate and develop based on the advantages, and continuously improved and optimized based on the disadvantages. The most important use of image processing is image recognition, which essentially involves identifying useful information in the image to be recognized through a trained network. Image recognition technology is widely studied due to its important application value and bright application prospects. Currently, image recognition is widely used in industries such as military, agriculture, and social life. In daily life, image recognition technology is constantly used: for example, in the popular smart home, character recognition is a relatively advanced image recognition. In addition, retinal scanning, fingerprint scanning, and other access control systems are also the same. Hospital clinical medical instruments use image recognition to judge and analyze the condition, and these applications have important practical significance [6].

2. Literature Review. With the progress of the times, the era of food and clothing has long passed. We now pay more attention to improving our quality of life, and science and technology are developing unprecedentedly. Whether it is dynamic publishing on social media or publishing articles on various forums, images

*Chongqing Industry & Trade Polytechnic, Chongqing, 408000, China (Corresponding author, HongmingPan8@163.com)

can add more colors and attract readers' attention. It can be seen that the use of images has penetrated into various fields; People in today's era are not only limited to dry and uninteresting written expressions, but also want to express their inner emotions and communicate with each other through images. Therefore, in the world of image transmission, image processing technology is particularly important. Li, Q. et al. proposed a novel and efficient neural convolutional network called MFU. The experimental results on various image denoising datasets (SIDD, DND, and synthetic Gaussian noise datasets) show that our MFU can produce comparable visual quality and accuracy results using state-of-the-art methods [7]. Vo, H. T. and others proposed a method based on deep learning network architecture for image classification of communication systems between autonomous vehicle [8]. Jun, M. et al. proposed a novel end-to-end dual stream convolutional neural network for single image dehazing. The network model consists of spatial information feature flow and high-level semantic feature flow. The spatial information feature flow preserves the detailed information of the dehazing image, while the advanced semantic feature flow extracts multi-scale structural features of the dehazing image. Designed a spatial information assistance module and placed it between feature flows. The experimental results show that the model proposed by the author can restore the desired visual effect without foggy images, and has good generalization performance in real haze scenes [9].

Artificial neural network technology has been successfully applied in multiple fields such as speech recognition, natural language processing, and computer vision. Convolutional neural networks are generated by simulating the visual nervous system and are mainly applied in the fields of natural language processing and computer vision. Convolutional neural networks are the intelligent crystallization of researchers continuously studying the features of advanced motor neurons. By referring to quantum theory and sharing feature parameters in a broad dimension, they significantly reduce the proportion of model storage. In the traditional sense, convolutional neural networks mainly extract some feature points first, and then represent the image through mathematical statistical models, which are often used to solve classification problems. The author applied convolutional neural networks to information image processing, and the final experiment proved that convolutional neural networks have very good performance in processing large-scale image datasets.

3. Method.

3.1. Convolutional Neural Network Structure. The most crucial aspect of convolutional neural networks in data processing is convolutional computation. The image data it processes is usually stored in BMP format, which is a format specified by the global non dynamic image storage organization. Its characteristic is to save the image to its original size without compression, so it requires a hardware system with large storage space. Therefore, after screening, the author selected the GE system, which is an advanced version of CPU and has fast pixel scanning ability, making it very suitable for image processing. And it has the ability to process multi format image conversion, and can transmit multi-dimensional images in multiple layers in parallel, laying a solid foundation for electronic information image processing [10].

3.2. Image Grayscale and Image Enhancement. Using high saturation red as the experimental background for image acquisition can improve the contrast between the parts and the background. The concept of image saturation comes from the HSV (Hue Saturation Value) color space, which is composed of hue h, saturation s, and brightness v. Due to the fact that images are generally represented by RGB color space components r, g, and b, conversion is required. Assuming $I_{max} = max\{r, g, b\}$ and $I_{min} = min\{r, g, b\}$, the relationship between the components s, v in the HSV color space and the RGB color space components r, g, and b is shown in equations 3.1 and 3.2:

$$s = \begin{cases} 0 \quad I_{min} \neq 0 \\ \frac{I_{max} - I_{min}}{T_{max}} \quad I_{max} \neq 0 \\ v = I_{max} \end{cases} \tag{3.1}$$

When there is reflection on the surface of the part, the color in that area appears bright white. The saturation value s in the reflection area of the part will decrease, and the contrast with the background will increase; When shadows appear in the background, the brightness of the area decreases, the saturation increases, and the contrast with the part also increases, as shown in the shaded part of the part. Therefore, extracting the image saturation layer as a grayscale image can effectively suppress the effects of reflection and shadows in the original RGB image. Next, the Luminance Contrast (LC) algorithm based on global contrast is adopted to further improve the contrast of the image [11]. The calculation method for the saliency value SalS (Ik) of pixel Ik in image I is shown in equation 3.3

$$SaIS(I_k) = \sum_{\forall I_i \in l} ||I_k - I_i|| \tag{3.2}$$

The range of values for I_i in the formula is [0.255], and $|| \cdot ||$ represents the distance between grayscale values.

3.3. Image binarization. The purpose of image binarization is to distinguish between parts and background pixels in the image. Firstly, the maximum inter class variance (OTSU) method is used for image binarization. Then, the morphological gradient operator is used to perform closure operations on the image, which can fill some small cracks and holes left by the binarization of the image, and fill the small broken curves into a whole [12].

3.4. Part area extraction. Using seed filling method to search for continuous white pixels in binary images can obtain the part regions and their boundary positions in the binary image. However, this algorithm takes a long time in the seed iteration process and has poor real-time performance. The author proposes a method based on image size transformation and adjusting boundary positions to improve the computational efficiency of seed filling method. The time required for seed filling method is closely related to the image size. Setting a size transformation parameter n and using bilinear interpolation to transform the image size, theoretically reducing the length and width of the image by n times before performing seed filling method operation, the time required will be shortened by 2n times, but it will bring significant operational errors. In order to reduce errors, the original image size is first reduced by n times to improve the computational speed of the seed filling method for obtaining the boundary position. Then, the reduced boundary position coordinates are enlarged by n times as the initial boundary position of the original size binary image. The coordinates of the pixels in the upper left and lower right corners of the initial boundary are (x_1, y_1), and (x_2, y_2), respectively, next, adjust the initial boundary in the original binary diagram to obtain the accurate part area boundary. Set the coordinates of the pixels in the upper left and lower right corners of the adjusted boundary to (x'_1, y'_1), (x'_2, y'_2), respectively.

The initial boundary position generally has a certain degree of error and needs to be adjusted. Each boundary has three positional relationships with the part: Intersection, separation, and tangent. The tangent state boundary is the precise boundary, and by traversing the pixel information of boundaries a, b, c, and d, the positional relationship between the boundary and the part area can be determined. If the boundary intersects with the part, it moves towards the outer direction. If it is apart, it moves towards the center of the image. Iterative operation is performed, and the change in position relationship is used as the termination condition to adjust the boundary to a tangent state, represented by (x'_1, y'_1), (x'_2, y'_2). In order to facilitate the training of subsequent CNN models, a rectangular image, namely the part area image, is extracted from the original RGB image based on the coordinates of boundary pixel points (x'_1, y'_1), (x'_2, y'_2). The short edges of the part area image are filled with red pixels to form a square, and then bilinear interpolation is used to unify the image size [13,14].

3.5. Overall Algorithm Structure. The part recognition algorithm proposed by the author first requires part area extraction, and then recognizes the extracted part area images, which can simplify the model and greatly reduce the workload of model training. The process of part recognition algorithm is shown in Figure 3.1.

3.6. CNN Based on Key Point Features of Parts. The training and testing of CNN based on part key point features are shown in Figure 3.2. Firstly, the SIFT algorithm is used to extract visual vocabulary vectors, namely feature vectors, from the image. These vectors represent locally invariant key point features in the image; Next, we use the Bag of Words (BoW) model to process the data, gather all feature vectors together, use K-means clustering algorithm to merge visual vocabulary with similar meanings, and construct a dictionary

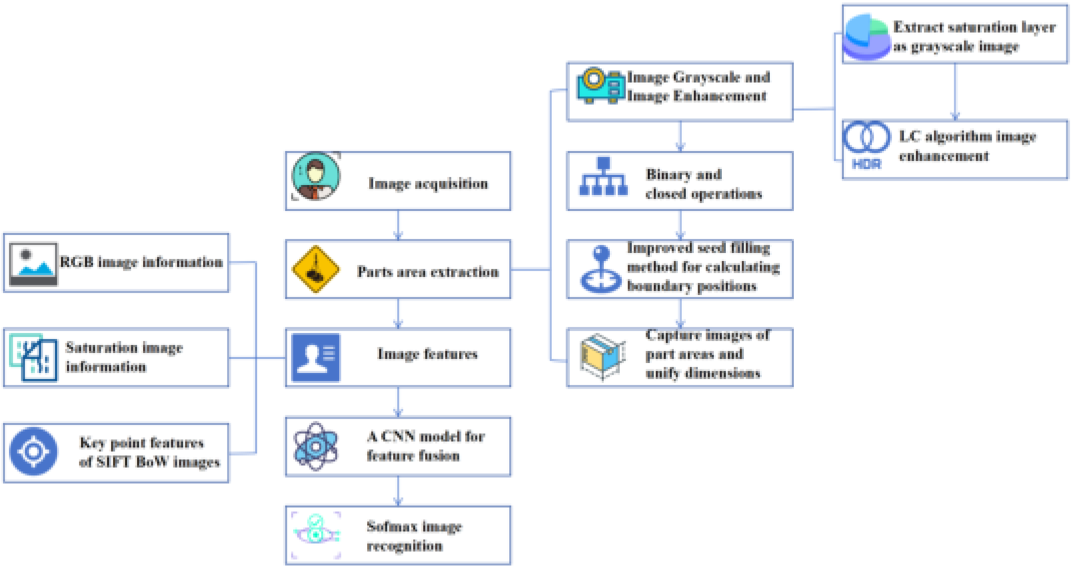

Extract saturation layer as grayscale image
LC algorithm image enhancement
HDR
Image Grayscale and Image Enhancement
Binary and closed operations
Improved seed filling method for calculating boundary positions
Capture images of part areas and unify dimensions
Image acquisition
Parts area extraction
Image features
A CNN model for feature fusion
Sofmax image recognition
RGB image information
Saturation image information
Key point features of SIFT BoW images

Fig. 3.1: Process of Part Recognition Algorithm

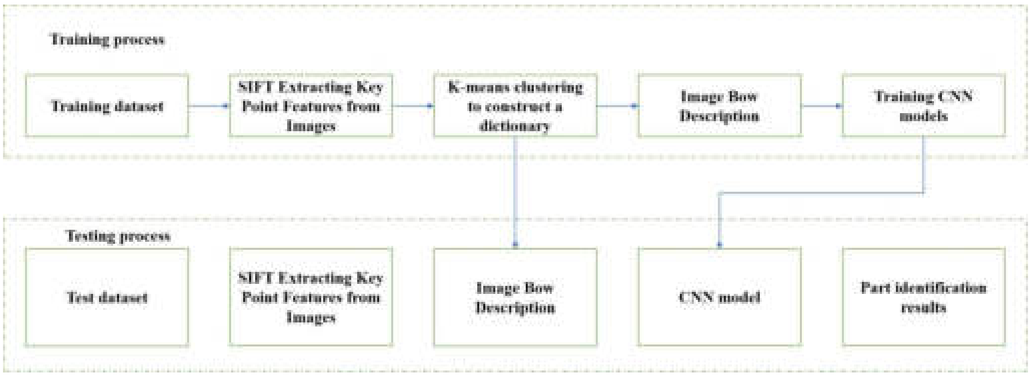

Training process
Training dataset
SIFT Extracting Key Point Features from Images
K-means clustering to construct a dictionary
Image Bow Description
Training CNN models
Testing process
Test dataset
SIFT Extracting Key Point Features from Images
Image Bow Description
CNN model
Part identification results

Fig. 3.2: Training and Testing of CNN Based on Key Point Features of Parts

containing K words, by counting the number of times each word in the dictionary appears in the image, the image can be represented as a K-dimensional feature vector, which is the BoW description of the image; Finally, the vector is used as input for a CNN based on part keypoints, and convolution and normalization operations are performed on it to obtain the feature vector of part keypoints [15].

3.7. CNN model for feature fusion. After experimental comparison, the VGG16 model was selected as the feature extraction network. The input was a 64 * 64 * 3 RGB image and a 64 * 64 * 1 saturation image stacked to form a 64 * 64 * 4 matrix, which underwent five double convolution operations and four max pooling operations. Due to the important role of the geometric shape of parts in part recognition, its importance is higher than features such as color and surface texture. By stacking high contrast saturation images with RGB images as inputs to the CNN model, the geometric shape features of the parts are enhanced during network training.

The SIFT-BOW model can extract local keypoint features of part images, which are feature vectors with a size of 1 × 1 × K. Firstly, K 1xK convolution operations and normalization operations are performed on them, and then enter the fully connected layer to obtain a 128 dimensional feature vector of part keypoints. The

Table 4.1: Part Area Extraction Time

number	Part Name	Time/s for different size transformation parameters n					
		n=1	n=2	n=4	n=6	n=8	n=10
01	Gear shaft	1.045	0.388	0.213	0.175	0.164	0.162
03	bearing	1.001	0.354	0.182	0.153	0.141	0.134
04	Driven shaft	1.076	0.386	0.222	0.182	0.164	0.158
05	bolt	1.003	0.347	0.187	0.153	0.148	-
07	Gearbox cover	1.535	0.621	0.383	0.336	0.326	0.316
08	Gearbox seat	1.46	0.538	0.276	0.233	0.213	0.206
09	Big gear	1.205	0.436	0.243	0.193	0.182	0.177
12	Safety valve box	1.624	0.437	0.241	0.202	0.190	0.183
15	Belt pulley	1.113	0.392	0.210	0.176	0.163	0.160

key points of the part are concatenated with the feature vectors of the part shape to obtain a 256 dimensional feature vector. After passing through a fully connected layer, its size is reduced to 1 × 1 × c. Finally, the recognition result is obtained using the Softmax algorithm. By concatenating local keypoint feature vectors at the end of the network, the feature loss generated during downsampling can be compensated for, the feature extraction ability can be enhanced, and the recognition accuracy can be improved [16].

3.8. Construction of experimental platform. Experimental environment: Windows 10, 64 bit operating system, memory size AMD Ryzen, GPU GTX-1650, and commonly used environments such as Python 3.6, Tensorflow 2.0, and Opencv were built. Place the tested part on a red background stage, use a strip LED light source for illumination, and keep the phone and computer in the same network environment using an IP camera for shooting and data transmission. In the experiment, the camera and desktop were kept horizontal, and the camera was about 36cm higher than the desktop. Collect samples under ambient light of different brightness levels in the experiment; The parts are arranged in a disorderly manner but not stacked on top of each other [17,18].

4. Results and Discussion. The dataset images for the experiment include 19 types of actual captured plunger pumps, gearboxes, gear oil pumps, safety valves, and ball valve components. 80 samples were collected for each type of component as initial samples. In order to increase data diversity, the images were expanded through horizontal and vertical flipping, as well as random changes in contrast and brightness, to improve the robustness of the model. A total of 1280 samples were obtained for each type of component. Train the model using a ratio of 7:3 between the training and testing sets. The original image size is 480 x 640, and the sample image size after extracting the part area is uniformly 64 x 64. Perform image grayscale, binarization, and morphological processing on the collected images in sequence, and then extract the part area. The core of the extraction algorithm is the size transformation parameter n. The author conducted experimental comparisons on the unchanged size of n=1 and the improved algorithm with n=2, 4, 6, 8, and 10, respectively, to explore the impact of different size transformation parameters n on the efficiency of part region extraction. The extraction time for some part areas collected in the experiment is shown in Table 4.1.

As shown in Table 4.1, directly extracting the part area from the original image, that is n=1, takes more than 1 second, which is inefficient. By changing the size of the image, the computational efficiency can be significantly improved. As n continues to increase, the improvement in computational efficiency gradually decreases. This is because image preprocessing takes time independent of the n value before size transformation. At the same time, in order to explore the impact of different size transformation parameters n on the quality of part region extraction, boundary position errors were calculated for some parts with n=2, 4, 6, 8, and 10 pairs, respectively. The results are shown in Table 4.2.

According to Table 4.2, when n=2, n=4, and n=6, the boundary position error pixels of each part are all lower than 3 pixels, indicating high accuracy; When n=8 and n=10, the boundary position error of the bolt significantly increases or even misses, due to the excessive reduction in image size causing loss of pixels in small parts. Therefore, considering the time required for extracting part regions in Table 4.3, the size transformation

Table 4.2: Boundary position errors for different parameters n

number	Part Name	Time/s for different size transformation parameters n				
		n=2	n=4	n=6	n=8	n=10
01	Gear shaft	1.24	1.24	1.24	1.3	1.34
03	bearing	0	0	0	0	0
04	Driven shaft	0.92	0.92	0.92	0.92	0.92
05	Snail inspection	1.6	1.6	2.6	14.8	-
07	Gearbox cover	1.04	1.04	1.04	1.04	1.04
08	Gearbox seat	0.42	0.42	0.42	0.42	0.42
15	Belt pulley	0.1	0.1	0.1	0.1	0.1

Table 4.3: Main parameters of CNN model

Parameter Name	Method/Value	Parameter Name	Method/Value
Number of images loaded each time	32	Training cycle	8
optimization algorithm	Adam	Single cycle time/min	8
Convolutional kernel	[3 × 3]	Trainable parameters/piece	2,290,724
Cluster center k	100	Number/type of sample categories	19

Table 4.4: Recognition Results of Different Algorithms

method	accuracy
PCA dimensionality reduction SVM	82.52%
SIFT-SVM	92.85%
SIFT-CNN	96.15%
CNN (direct extraction of image features)	97.80%
Feature fusion CNN model	98.84%

parameter n=4 is the optimal parameter for the part region extraction algorithm [19].

The selection of n value needs to consider the size of the parts. If identifying smaller parts, such as bolts, nuts, etc., the n value should not be too large. A value between 2 and 6 can maintain high recognition accuracy. If the size of the identified parts is large, such as large gears, safety valve bodies, etc., taking an n value between 6 and 10 can bring high efficiency and ensure recognition accuracy. The training time of this algorithm is only 8 minutes, which requires several hours of training compared to large object detection models. It also avoids the annotation work on the dataset. The training time of this algorithm is short and practical. After experimental comparison, the CNN model adopts the following parameters, as shown in Table 4.3.

In order to verify the progressiveness and effectiveness of the feature fusion CNN model, it is compared with a single model and existing recognition methods, and the results are shown in Table 4.4.

According to Table 4.4, the accuracy of the model using only Principal Component Analysis (PCA) and SIFT for feature extraction is not high, while the CNN model is good at extracting features from images and generalization. For CNN models that do not contain SIFT features, the accuracy of directly extracting features from RGB images is 97.80%. On this basis, the feature fusion CNN model superimposes saturation images on the input layer, enhancing the shape features of parts in network training. Combined with the key point feature vectors concatenated at the end of the network, the accuracy of part image recognition reaches 98.84%. Due to the unified size of 64 × 64 during the part extraction process, the CNN model takes about 2ms to predict, and the overall recognition algorithm speed is about 5fps, including the time for part area extraction.

5. Conclusion. The author proposes a research on electronic information image processing based on convolutional neural networks. In order to address the issues of easy reflection, unevenness, and uneven lighting on the surface of mechanical parts, as well as the interference caused by shadows and mixed other parts on recognition, a seed filling method is proposed to extract part regions through image saturation transformation and size transformation. This method can quickly and accurately extract part regions from images. Dividing the task of part image recognition into two steps: part region extraction and image recognition is beneficial for improving the scalability of the algorithm. Only by retraining a lightweight CNN classification model, new parts can be added for recognition, avoiding tedious part position annotation and complex boundary regression operations, which can reduce the computational power requirements of computing equipment. The author proposes a feature fusion CNN model that overlays RGB images with saturation images and inputs them into the CNN model. At the end of the CNN, feature vectors based on key points in the part image are concatenated and then passed down through the network. This can compensate for the loss of local information in the image during convolutional pooling, thereby further improving the accuracy of the recognition algorithm.

REFERENCES

[1] Yanshan, L. I., Zhou, L., Fan, X. U., & Chen, S. (2022). Ogsrn:optical-guided super-resolution network for sar image, 35(5), 204-219.

[2] He, H., Yang, P. F., Zhang, P. F., Li, G., & Zhang, T. C. (2023). Single-photon source with sub-mhz linewidth for cesium-based quantum information processing, 18(6), 61303.

[3] Zhang, Y. L., Feng, X. L., Chang, X. L., & Tie, L. M. (2023). Impacts of canopy structure on the sub-canopy solar radiation under a deciduous forest based on fisheye photographs, 15(3), 150-160.

[4] Wang, Z., Li, L., Xing, L., Wang, J., Sun, K., & Ma, H. (2023). Information purification network for remote sensing image super-resolution. Tsinghua Science and Technology, 28(2), 310-321.

[5] Liu, X., Nicolau, J. L., Law, R., & Li, C. (2023). Applying image recognition techniques to visual information mining in hospitality and tourism. International Journal of Contemporary Hospitality Management, 35(6), 2005-2016.

[6] Zhang, D., Zhao, L., Duanqing, X. U., & Dongming, L. U. (2022). Dual-constraint burst image denoising method. Frontiers of Information Technology & Electronic Engineering, 23(2), 220-233.

[7] Guo, X., He, P., Lv, X., Ren, X., Li, Y., & Liu, Y., et al. (2023). Material decomposition of spectral ct images via attention-based global convolutional generative adversarial network, 34(3), 143-153.

[8] Vo, H. T., Nguyen, N. L., Ngoc, D. N., Do, T. H., & Pham, Q. D. (2023). Binary Image Classification Using Convolutional Neural Network forV2V Communication Systems, 28(4), 743-753.

[9] Jun, M., Yuanyuan, L., Huahua, L., & You, M. (2022). Single-image dehazing based on two-stream convolutional neural network. Journal of Artificial Intelligence Technology (English), 76(9), 3859-3876.

[10] Yang, W., Huang, Z., & Zhu, W. (2023). A first-order rician denoising and deblurring model. Inverse Problems and Imaging, 17(6), 1139-1164.

[11] Chen, Q. (2022). Research on 3d mfl testing of wire rope based on empirical wavelet transform and srcnn. Journal of Vibroengineering, 24(4), 14.

[12] Singh, B. K., Nair, N., Falgun, P. A., & Jain, P. (2022). Quantitative evaluation of denoising techniques of lung computed tomography images: an experimental investigation. International Journal of Biomedical Engineering and Technology, 38(2), 151.

[13] Wen, Y., Sun, J., & Guo, Z. (2022). A new anisotropic fourth-order diffusion equation model based on image features for image denoising. Inverse Problems and Imaging, 16(4), 895-924.

[14] Joshi, S., Singla, N., Ahuja, S., Saini, S. K., Thakur, N., & Jindal, P., et al. (2022). Denoising of computed tomography using bilateral median based autoencoder network. International Journal of Imaging Systems and Technology, 32(3), 935-955.

[15] Huapeng Tang, Danyang, Q., Mengying, Y., Jiaqiang, Y., & Gengxin, Z. (2023). Research on color image matching method based on feature point compensation in dark light environment, 29(1), 78-86.

[16] Liu, X., Nicolau, J. L., Law, R., & Li, C. (2023). Applying image recognition techniques to visual information mining in hospitality and tourism. International Journal of Contemporary Hospitality Management, 35(6), 2005-2016.

[17] ZHOU Qian-fei,WANG Hui-fang,QING Guang-wei,HU Jing-bo. (2022). Detection of crane structural connecting bolts falling off based on uav image recognition. Manufacturing Automation, 44(12), 20-22.

[18] Liu, B., Li, J., Yang, X., Chen, F., Zhang, Y., & Li, H., et al. (2023). Diagnosis of primary clear cell carcinoma of the liver based on faster region-based convolutional neural network. Chinese Medical Journal, 136(22), 2706-2711.

[19] Zhao, S., Mao, G., Xiong, B., Huang, S., & Lin, J. (2023). Spatio-temporal graph mining model based on graph wavelet convolutional neural network. Computer Engineering, 49(7), 85-93.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1300–1307, DOI 10.12694/scpe.v26i3.4210

MODELING OF SECURITY AND PRIVACY ARCHITECTURE FOR PROTECTING DATABASES IN CLOUD COMPUTING INFRASTRUCTURE

XIAOHUI ZHANG [*], SONGKUN JIAO[†], JUNFENG WANG[‡], AND CUILEI YANG[§]

Abstract. In order to prevent data leakage and ensure the security of tenant's private data, and to enable tenants to have a precise understanding of the security level of their private data, the author proposes a modeling of the security and privacy architecture for protecting databases in cloud computing infrastructure. The author proposes a document database privacy protection architecture, which builds upon the existing architecture by adding a privacy protection layer between the application layer and the storage layer, forming a new service deployment architecture. Then, the author introduced the basic methods of privacy protection based on facial document databases. In order to adapt to the data structure system based on document storage for document databases, the author designed a basic method of privacy protection for document databases based on segmentation and obfuscation. By utilizing the free nature of document oriented database patterns, privacy protection data can be achieved through appropriate segmentation. For nested document structures, the author designed a document structure tree to retain document structure information. The results show that by comparing the experimental data of the 50w and 100w groups horizontally, it can be found that under the same cutoff score, as the number of database data increases, the time for data queries also increases accordingly, the query time of database A has increased by nearly 300ms compared to database B, and the additional time of the other groups is also roughly the same. By vertically comparing the experimental data of the 50w and 100w groups, it can be found that the query time of the C database is nearly 200ms longer than that of the A database, and as the sharding factor increases, the query time also increases accordingly, but the proportion of increase begins to slow down. By comparing data with the same segmentation factor but different data volumes, it can be seen that the impact of data volume on query time is positively correlated. This model can ensure that the database system is transparent to users at the view layer after privacy protection, and ensure the correctness and integrity of privacy data.

Key words: Software as a Service, Data privacy protection, Targeting document databases, Data partitioning

1. Introduction. In this era, major enterprises strive to obtain valuable information from vast amounts of data by collecting data, and data has become an important information asset [1]. Enterprises can utilize this data again to enhance market competitiveness, making its value no longer singular. Therefore, data has become a valuable asset, important economic input, and cornerstone of new business models for companies [2]. However, data has become a challenge of this era due to its massive, high-speed, and diverse characteristics [3]. The rapid development of cloud computing technology provides technical support for big data processing, providing users with computing and data storage services. Currently, there are three service models for cloud computing: Infrastructure as a Service, IaaS, users rent the infrastructure of cloud platforms, such as computer hardware resources, in order to obtain the desired services; Software as a-S service, SaaS users can rent software resources and access them from browsers, application programming interfaces, etc., users only need to configure their own application software; Platform as a Service, PaaS: Users can rent pre built software platforms in the cloud, such as operating systems, software development environments, etc. The popularity of intelligent mobile devices has led to a large number of mobile applications generating massive amounts of unstructured data [4]. Meanwhile, the increasing popularity of Web 2.0 applications such as Google and Facebook also requires the analysis and storage of large amounts of unstructured data. Ordinary relational databases are no longer able to handle these unstructured data, and database systems are undergoing a revolution. Moreover, these

[*] Department of Rail Transit, Shijiazhang Institute of Railway Technology, Shijiazhuang, Hebei, 050000, China. (XiaohuiZhang9@126.com)

[†] China nuclear power engineering Co., Ltd.Hebei branch, Shijiazhuang, Hebei, 050000, China. (SongkunJiao@163.com)

[‡] Department of Rail Transit, Shijiazhang Institute of Railway Technology, Shijiazhuang, Hebei, 050000, China. (JunfengWang6@126.com)

[§] Department of Rail Transit, Shijiazhang Institute of Railway Technology, Shijiazhuang, Hebei, 050000, China. (Corresponding author, CuileiYang9@163.com)

applications no longer only focus on issues such as consistency assurance that relational databases excel at, but also pay more attention to performance, scalability, and so on [5]. Traditional relational models use standardized SQL queries and access them through common interfaces such as JDBC and JDBC. Despite the functionality implemented by each relationship system supplier, There are slight changes in both SQL and system interfaces, but relational database systems are relatively interchangeable due to their widespread acceptance of standards. Faced with the challenges of the big data era, traditional relational databases exhibit poor scalability and slow retrieval and read/write speeds when querying simultaneously. Although the application of database distributed storage and horizontal and vertical segmentation technologies can alleviate the above-mentioned problems in processing massive data. But data migration is difficult, with high performance requirements and relatively high management costs [6].

2. Literature Review. Encryption is a traditional and effective method of protecting data privacy. The architecture proposed by Aldawibi, O. O., and others is aimed at addressing most privacy issues in cloud computing. The main idea behind the proposed architecture is to segment data and store it on many clouds using third parties [7]. Zhou, Z. et al. proposed a practical data auditing scheme with retrievability and indistinguishable privacy protection functions, which can effectively audit the status of outsourced data. Improved reversible Bloom filter (IBF) to locally compress redundancy, which can retrieve corrupted data without prior context. In addition, an indistinguishable privacy protection model has been defined to capture the complete semantics of duplicate audit attacks and achieve indistinguishability in auditing. It has been proven that the proposed scheme is secure against adaptive message selection attacks and indistinguishable privacy protection against duplicate audit attacks [8]. Ahmadi, S. and others believe that privacy protection cloud computing is an emerging technology with many applications in various fields. The reason why cloud computing is important is because it has scalability, adaptability, and improved security. Similarly, privacy in cloud computing is also important as it ensures the integrity of data stored in the cloud remains unchanged [9]. The author referred to the traditional methods of data segmentation and data obfuscation in relational databases, and designed a basic privacy protection method for document databases based on value segmentation and key obfuscation, taking advantage of the free nature of face to face document database patterns. By utilizing the free nature of document oriented database patterns, privacy protection data is achieved by appropriately segmenting and obfuscating the segmented keys. At the same time, the author combines the privacy protection architecture proposed in Chapter 2 to provide a basic operational model for a protected database system. This model can ensure that the database system is transparent to users at the view layer after privacy protection, and ensure the correctness and integrity of privacy data.

3. Research Methods.

3.1. Privacy Protection System Architecture. At present, the SaaS system uses a document database oriented architecture with a two-layer structure of application server and database server. The application and database interact directly, and the application can access the data in the database, Internal staff of SaaS service providers can also see the data stored by tenants in the database [10-11]. The data capture of tenants is easily obtained by internal employees of SaaS service providers, posing a data security risk. Therefore, the author's privacy protection architecture adds a privacy policy layer between the application layer and the data storage layer, as shown in Figure 3.1. The privacy protection architecture proposed by the author is divided into three layers, including three types of servers: Application Server, Privacy Policy Server, and Storage Server [12]. The application server is responsible for providing SaaS application services to tenants. The application server is a traditional SaaS application that only needs to submit data operation requests to the underlying layer, without worrying about whether tenant data needs to execute privacy protection policies or the execution methods of privacy protection policies; The privacy policy server is responsible for storing the privacy protection policies of tenants, verifying their identity after submitting operation requests, and implementing privacy protection for their data and operations based on their privacy protection policies. The privacy policy server includes an application interface module, identity authentication module, policy module, and conversion module. The storage server is responsible for storing tenant data and accepting data operation requests from the upper layer.

In this architecture, the deployment of privacy policy servers by SaaS service providers has no impact

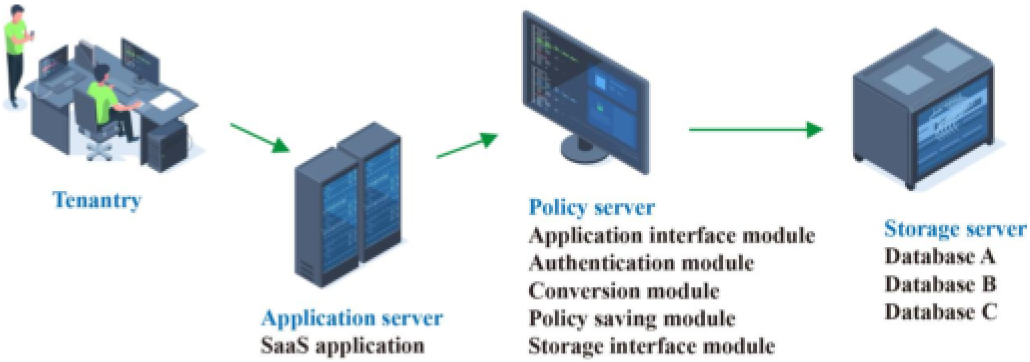


Fig. 3.1: Privacy Protection Architecture Diagram

on the execution of SaaS applications. When tenants need to manipulate data, they only need to send their identity information and operation requests to the SaaS application server, the SaaS application server executes application logic, sends tenant identity information and database operation statements to the next layer when data access is required, and waits for the return result. If a privacy policy server is deployed, the privacy policy server unlocks tenant policy information through tenant identity information and performs further operations [13]. If a privacy policy server is not deployed, the operation command will be directly transmitted to the storage server. This design solves the problem of tenant privacy protection policies and SaaS applications.

Privacy policy servers and storage servers are two independent server clusters. Tenant privacy policies are encrypted and stored in the privacy policy server. Even if SaaS service providers view tenant data in the storage server, they cannot obtain tenant privacy information because they cannot access the tenant privacy policies on the privacy policy server.

Furthermore, privacy protection servers can be provided as independent services, managed by specialized service providers. The service provider that provides privacy protection saves the tenant's privacy protection policy, but does not have tenant data. SaaS application service providers have tenant data, but cannot restore tenant data [14]. By isolating privacy policies from tenant data, we ensure the security of tenant privacy.

3.2. Basic methods of privacy protection. As is well known, the efficiency of querying encrypted data is low, so the author adopts a data segmentation method. Data segmentation is a method of protecting privacy data in relational databases, which effectively balances data read and write performance with data integrity [15]. All document data in document databases is composed of Key Value pairs, and the data of each Key Value pair is the minimum granularity for privacy data protection in document databases. If it can be guaranteed that attackers cannot obtain privacy information from Key Value pairs after obtaining data, then this privacy protection method is an available privacy protection method.

Definition 1. Value segmentation and key confusion Value segmentation is the process of segmenting the Value values of attributes that need to be protected in order to protect privacy data. $\{K : V\} \rightarrow \{K : [V_{r_1}, V_{r_2}, \cdots, V_{r_n}]\}$ While segmenting values, we will also segment each V_{r_i}; Renaming $K \rightarrow \{K_{r_1}, K_{r_2}, \cdots, K_{r_n}\}$ using the Key. In the process of renaming keys, we only ensure that each key can correspond to a fixed value, that is, there is a constraint $K_{r_i} \rightarrow V_{r_i}$ that makes $\{K : V\} \rightarrow \{\{K_{r_1} : V_{r_1}\}, \{K_{r_2} : V_{r_2}\}, \cdots, \{K_{r_n} : V_{r_n}\}\}$, while the order of keys is confidential. We refer to the renaming process as key obfuscation. The arrangement order of the confused $\{K_{r_1}, K_{r_2}, \cdots, K_{r_n}\}$ is the privacy protection strategy we obtained [16].

This segmentation divides the S SN attribute into three independent key value pairs, and the privacy protection policy after segmentation is $K_{SSN}[dd, ff, ss]$. By hiding the combination order of these three key value pairs, the purpose of protecting tenant privacy data is achieved.

Pattern freedom is a major feature of document oriented databases, and the author's approach is to utilize the feature of pattern freedom in document oriented databases to segment private data. Pattern freedom means that the length of certain attributes is not fixed [17]. We assume that the attributes to be protected for privacy are within a range, meaning that each attribute will have a minimum length and a maximum length. When formulating privacy policies, we determine the number of obfuscated keys based on the minimum length of attributes. And those big data with variable lengths are not within the author's scope of consideration.

When performing value sharding, it is necessary to determine the number of shards for each attribute. The size of the number of shards indirectly determines the degree of privacy protection and also affects the performance of the server in processing private data. The author names this number of shards as the sharding factor. The more partitions there are, the higher the level of privacy protection, but the greater the workload of reorganizing values. If the number of partitions is exactly equal to the length of the private data, it can be considered that a special encryption algorithm has been used for the data. The author determines the segmentation factor for each attribute based on the maximum length of the attribute, which ensures that each segmented attribute block has the same length. This method allows certain keys to have empty values and does not store keys with empty values.

Below is a segmentation algorithm for the basic methods of privacy protection in document oriented databases, Value segmentation can be divided into two types: fixed length segmentation and fixed number of blocks segmentation. Fixed length segmentation refers to the fixed length of the segmented blocks for a certain attribute. Fixed block segmentation refers to the fixed length segmentation used for a certain attribute, where the number of blocks generated is fixed. We use fixed length segmentation for privacy attributes, and the algorithm's parameter pl is the length of each block.

3.3. Nested Document Privacy Protection Methods. Nested document refers to treating the entire document as a key value in another document. Nested document structure is an important feature of document oriented databases, which allows data to be organized more naturally without having to be stored in a flat structure. We need to protect privacy data in object-oriented document databases, and we must fully consider how to preserve the structural information of nested documents while protecting privacy.

In order to protect the data in the embedded document, we must understand all the information about the document structure. Therefore, we generated a document structure tree, as shown in Figure 3.2. The root node of the tree is a data table for privacy protection, and the hierarchical relationship of the tree is established based on the nested relationship of documents. The author assumes that all documents in each data table contain the same structure [18]. Nodes containing value ultimately become leaf nodes, and each leaf node has a leaf node as its successor. For each leaf node, there are three attributes: Type, minimum length, and maximum length. In the diagram, type 0 represents numbers and type 1 represents characters. As shown in Figure 3, the age attribute is of numerical type, with a minimum length of 1 and a maximum length of 3. The maximum length of an attribute is used to determine the number of renamed keys in case of key obfuscation, while the minimum length of an attribute determines the maximum value of the splitting factor.

Through the document structure tree, the domain names of nested structure attributes can be obtained, which represent the structural information of attributes in the nested document. For example, if the domain name of the attribute phone is address.tel, we can obtain the full name of the attribute address.tel.phone. When we need to segment a certain attribute, the privacy policy records the full name of the attribute in order to preserve the nested information of the document in future transformations and combinations, and anonymize it in the document structure tree.

When segmenting a certain attribute in a document, first find the node of that attribute in the document structure tree, and determine the number of segmentation blocks based on the minimum length of the attribute. After the segmentation is completed, the obfuscated blocks will be stored under the root node, and the node names occupied by this attribute will be anonymized in the document structure tree, while the node attributes will be retained. For example, by segmenting the addres.zip attribute, the privacy policy after segmentation is set to $K_{address.zip}[z1, z3, z2]$.

The anonymization process of document structure tree is irreversible, and the confidence of document location after familiarizing oneself with anonymity will be saved in the privacy policy. If the application needs to view document structure information, the document structure tree can be reconstructed according to the privacy policy [19].

3.4. Privacy data operation methods. This section introduces the processing process of data operation requests on the policy server when tenants perform addition, deletion, modification, and query operations.

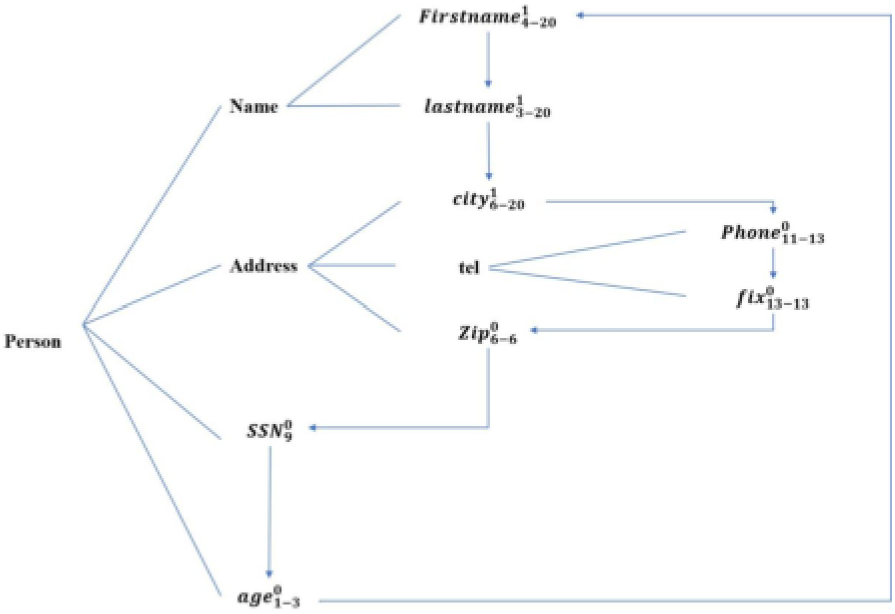


Fig. 3.2: Document Structure Tree

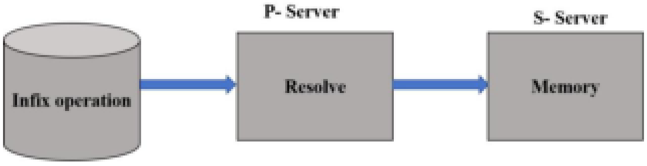


Fig. 3.3: Insertion Operation Process

When a tenant wants to store data, they submit a request to the policy server, which includes the tenant authentication information ID (T), tenant key PK, and storage operation request $op(\langle K_a, V_a >< K_\beta, V_\beta, V_\beta >< K_\theta, V_\theta \rangle)$. The policy server decrypts the request through the decryption function $D_{(PK.K)}$; Obtain the tenant's privacy policy $K_a[K_{\alpha r1}, K_{cr2} \cdots], K_p[K_{\rho r1}, K_{\rho r2} \cdots]$, then segment the tenant's data based on the tenant's privacy policy, pair the segmented data with the obfuscated Key to form $\langle K_{\alpha r1}, V_{r1} >< K_{\alpha r2}, V_{r2} \rangle \cdots, \langle K_{\beta r1}, V_{r1} >< K_{\beta r2}, V_{r2} \rangle \cdots, \langle K_\theta, V_\theta \rangle$, and finally forward the processed data to the storage server. After receiving the data sent by the policy server, the storage server stores the data (Figure 3.3).

When a tenant wants to search, update, and delete data, they submit their authentication information ID (T), tenant key PK, and data operation request $op(K_a = V_a, K_\beta = V_\beta)$ to the policy server. After decrypting the privacy policy, the policy server converts the request into $op(K_{\alpha r1} = V_{r1} \& K_{\alpha r2} = V_{r2} \& \cdots, K_{\beta r1} = V_{r1} \& K_{\beta r2} = V_{r2} \& \cdots)$ based on the privacy protection policy $K_a[K_{\alpha r1}, K_{cr2} \cdots], K_p[K_{\rho r1}, K_{\rho r2} \cdots]$. Then send the converted request to the storage server while waiting for the return information from the storage server [20]. After receiving a tenant request, the storage server processes it based on the tenant request. If it is a query request, the query result is returned to the policy server. The policy server recombines the query results based on the privacy policy and forwards them to the tenant.

When a tenant wants to update or delete data, they submit their authentication information ID (T), tenant key PK, and data operation request $op(K_a = V_a, K_\beta = V_\beta)$ to the policy server. After decrypting the privacy policy, the policy server converts the request into $op(K_{\alpha r1} = V_{r1} \& K_{\alpha r2} = V_{r2} \& \cdots, K_{\beta r1} = V_{r1} \& K_{\beta r2} = V_{r2} \& \cdots)$ based on the privacy protection policy $K_a[K_{\alpha r1}, K_{cr2} \cdots], K_p[K_{\rho r1}, K_{\rho r2} \cdots]$, then send

Table 4.1: Experimental Configuration Table

database	Data volume (10000)	Splitting factor
A	25	0
B	50	0
C	100	0
D	25	4
E	50	4
F	100	4
G	25	8
H	50	8
I	100	8

the converted request to the storage server while waiting for the return information from the storage server. After receiving a tenant request, the storage server processes it according to the tenant request. If it is a deletion operation, the storage server returns confirmation information to the policy server, which forwards the confirmation information to the tenant; If it is an update operation, return the updated data or confirmation information based on the tenant configuration.

4. Result analysis. The author designed a basic method for privacy protection in document databases based on val ue segmentation and key obfuscation. In order to maintain the nested document feature of document oriented databases, the author designed a document structure number [21]. By utilizing the free nature of document oriented database patterns, privacy protection data is achieved by appropriately segmenting and obfuscating the segmented keys. The author's experiment used a PC as the database server, with the server configuration as follows:

CPU: 4 cores Intel(R) Core(R) i5-2400 3.IOGHz
Memory: 2GB
Hard disk: 500GB
Operating System: Red Hat Enterprise Linux&. 2 32bit
Database system: MongoDB 2.4.9

The author's experiment used a power system electricity meter as the data mode. Relevant datasets were generated through simulation. The sharding factor is the number of sharding blocks for tenant privacy data. This experiment will test the impact of the size of the sharding factor on database performance and provide reference for determining the sharding factor. The experiment will establish 6 databases, and the data volume of each database is shown in Table 1. The same amount of data stored in the database for this experiment is the same, that is, database A, D. G stores the same copy of data, database B, E. H stores the same copy of data, database C, F. I stores identical copies of data. The D-I database will segment the data date attribute, and the segmentation factors are shown in Table 4.1.

The experiment will perform the same query operation on 6 databases, query the data in the data date column, and check the query time. The column has not been indexed, and the data queried for the same data copy in this experiment is the same, data date The experimental results are shown in Figure 4.1. This experiment has made every effort to eliminate the bias caused by database caching during the experimental process [22,23].

By comparing the experimental data of the 50w and 100w groups horizontally, it can be found that under the same cutoff score, as the number of database data increases, the time for data queries also increases accordingly, the query time of database A has increased by nearly 300ms compared to database B, and the additional time of other groups is also roughly the same [24]. By vertically comparing the experimental data of the 50w and 100w groups, it can be found that the query time of the C database is nearly 200ms longer than that of the A database, and as the sharding factor increases, the query time also increases accordingly, but the proportion of increase begins to slow down. By comparing data with the same segmentation factor but different data

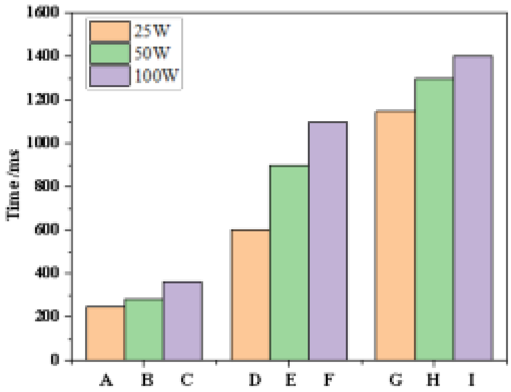


Fig. 4.1: The impact of segmentation factors on efficiency

volumes, it can be seen that the impact of data volume on query time is positively correlated [25]. Through this experiment, it can be seen that the size of the sharding factor has an impact on the search performance of a data. The larger the sharding factor, the longer the additional time required for data queries. Therefore, when setting the sharding factor, it is necessary to fully consider the performance requirements of tenants. At the same time, we also found that although data partitioning affects the query speed of data, this impact is not as significant as increasing the size of the database. The impact of data partitioning on the query performance of the database does not exceed 30%. This is still done under the premise of eliminating the impact of database performance optimization through memory. If the database is allowed to perform query optimization, the impact of this partitioning will be even lower. Therefore, the experiment in this section demonstrates the feasibility of using the privacy protection method proposed by the author in document oriented databases.

5. Conclusion. In order to adapt to the storage characteristics of document oriented databases, the author designed a basic privacy protection method for document oriented databases based on value segmentation and key obfuscation. In order to maintain the feature of nested documents in document oriented databases, the author designed a document structure tree. By utilizing the free nature of document oriented database patterns, privacy protection data is achieved by appropriately segmenting and obfuscating the segmented keys. In order to ensure the security of tenant privacy policies, the author provides a basic operation model for a protected database system. This model can ensure that the database system is transparent to users at the view layer after privacy protection, and ensure the correctness and integrity of privacy data.

REFERENCES

[1] Abba Ari, A. A., Ngangmo, O. K., Titouna, C., Thiare, O., Mohamadou, A., & Gueroui, A. M. (2024). Enabling privacy and security in Cloud of Things: Architecture, applications, security & privacy challenges. Applied Computing and Informatics, 20(1/2), 119-141.

[2] Saini, D. K., Kumar, K., & Gupta, P. (2022). Security issues in IoT and cloud computing service models with suggested solutions. Security and Communication Networks, 2022, 1-9.

[3] Bibal Benifa, J. V., & Venifa Mini, G. (2020). Privacy based data publishing model for cloud computing environment. Wireless Personal Communications, 113, 2215-2241.

[4] El Kafhali, S., El Mir, I., & Hanini, M. (2022). Security threats, defense mechanisms, challenges, and future directions in cloud computing. Archives of Computational Methods in Engineering, 29(1), 223-246.

[5] Sheikh, M. S., Liang, J., & Wang, W. (2020). Security and privacy in vehicular ad hoc network and vehicle cloud computing: a survey. Wireless Communications and Mobile Computing, 2020, 1-25.

[6] Gill, S. H., Razzaq, M. A., Ahmad, M., Almansour, F. M., Haq, I. U., Jhanjhi, N. Z., ... & Masud, M. (2022). Security

and privacy aspects of cloud computing: a smart campus case study. Intelligent Automation & Soft Computing, 31(1), 117-128.

[7] Aldawibi, O. O., Sharf, M. A., & Obaid, M. M. (2022). Cloud computing privacy: concept, issues and solutions. 2022 IEEE Symposium on Industrial Electronics & Applications (ISIEA), 33(7), 810-819.

[8] Zhou, Z., Luo, X., Wang, Y., Mao, J., Luo, F., & Bai, Y., et al. (2023). A practical data audit scheme with retrievability and indistinguishable privacy-preserving for vehicular cloud computing. IEEE Transactions on Vehicular Technology, 117(1), 109-127.

[9] Ahmadi, S., & Salehfar, M. (2022). Privacy-preserving cloud computing: ecosystem, life cycle, layered architecture and future roadmap, 34(6), 3121-3135.

[10] Lee, C., & Ahmed, G. (2021). Improving IoT privacy, data protection and security concerns. International Journal of Technology, Innovation and Management (IJTIM), 1(1), 18-33.

[11] Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2021). Exploring motivations for online privacy protection behavior: Insights from panel data. Communication Research, 48(7), 953-977.

[12] Wu, Y., Song, L., & Liu, L. (2021). The new method of sensor data privacy protection for IoT. Shock and Vibration, 2021, 1-11.

[13] Ren, Y., Liu, W., Liu, A., Wang, T., & Li, A. (2022). A privacy-protected intelligent crowdsourcing application of IoT based on the reinforcement learning. Future generation computer systems, 127, 56-69.

[14] Humayun, M., Jhanjhi, N. Z., Alruwaili, M., Amalathas, S. S., Balasubramanian, V., & Selvaraj, B. (2020). Privacy protection and energy optimization for 5G-aided industrial Internet of Things. IEEE Access, 8, 183665-183677.

[15] Zhang, Q., Li, Y., Wang, R., Liu, L., Tan, Y. A., & Hu, J. (2021). Data security sharing model based on privacy protection for blockchain-enabled industrial Internet of Things. International Journal of Intelligent Systems, 36(1), 94-111.

[16] Akter, M., Moustafa, N., Lynar, T., & Razzak, I. (2022). Edge intelligence: Federated learning-based privacy protection framework for smart healthcare systems. IEEE Journal of Biomedical and Health Informatics, 26(12), 5805-5816.

[17] Sun, Z., Wang, Y., Cai, Z., Liu, T., Tong, X., & Jiang, N. (2021). A two-stage privacy protection mechanism based on blockchain in mobile crowdsourcing. International Journal of Intelligent Systems, 36(5), 2058-2080.

[18] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., ... & He, B. (2021). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3347-3366.

[19] Diaz-Ordoñez, M., Rodríguez Baena, D. S., & Yun-Casalilla, B. (2023). A new approach for the construction of historical databases—NoSQL Document-oriented databases: the example of AtlantoCracies. Digital Scholarship in the Humanities, 38(3), 1014-1032.

[20] Istiqamah, A. N., & Wiharja, K. R. S. (2021). A schema extraction of document-oriented database for data warehouse. International Journal on Information and Communication Technology (IJoICT), 7(2), 36-47.

[21] Maicha, M. E., Ouinten, Y., & Ziani, B. (2022). UML4NoSQL: A Novel Approach for Modeling NoSQL Document-Oriented Databases Based on UML. Computing and Informatics, 41(3), 813-833.

[22] Messaoud, I. B., Alshdadi, A. A., & Feki, J. (2021). Building a document-oriented warehouse using NoSQL. International Journal of Operations Research and Information Systems (IJORIS), 12(2), 33-54.

[23] Gusarenko, A. S. (2022). Situation-oriented databases: processing heterogeneous documents of microservices in a document-based storage. Modelirovanie, Optimizatsiya i Informatsionnye Tekhnologii= Modeling, Optimization and Information Technology, 10(4), 1-16.

[24] Davardoost, F., Sangar, A. B., & Majidzadeh, K. (2020). Extracting OLAP cubes from document-oriented NoSQL database based on parallel similarity algorithms. Canadian Journal of Electrical and Computer Engineering, 43(2), 111-118.

[25] Maté, A., Peral, J., Trujillo, J., Blanco, C., García-Saiz, D., & Fernández-Medina, E. (2021). Improving security in NoSQL document databases through model-driven modernization. Knowledge and Information Systems, 63, 2209-2230.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 7, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1308–1316, DOI 10.12694/scpe.v26i3.4211

DESIGN OF POWER GATEWAY BASED ON EDGE COMPUTING AND RESEARCH ON DATA TRANSMISSION SECURITY

YAJIE LI[*], TAO MING[†], JIANGTAO GUO [‡], YUAN CAO [§] AND HONG LI [¶]

Abstract. In order to realize the centralized processing and remote management of data in the power grid, the author proposes the design of power gateway based on edge computing and the research on data transmission security. The author builds a computing service model in the gateway through Docker virtualization technology, analyzes the application of edge computing in the power industry, builds services based on Python language and Docker technology, and designs the edge computing gateway system. The experimental results indicate that: This model can accurately obtain the coordinate position and rotation angle of the grounding knife, and selects a convolution kernel size of 6 × 6 based on the information loss rate and defect removal rate of gateway image processing. Zone A is in the closed state (92 °, 98 °), zone B is in the middle state (98 °, 175 °), and zone C distinguishes the closed state (175 °, 179 °). This recognition algorithm can accurately analyze the status of the grounding knife, so it can be used to build a power gateway to ensure information security. The problem of logic confusion in Docker's design of power gateway is solved, follow up work is carried out with this idea, and a power gateway based on edge computing is designed and implemented.

Key words: Edge computing, Power gateway, Docker, cloud computing

1. Introduction. The concept of smart grid has emerged and gradually become a hot topic for research and exploration in the global power industry. With the development of society and economy, as well as technological progress, human society's dependence on electricity is increasing.

The intelligence network, also known as the intelligence of the power grid, is built on the basis of an integrated, high-speed bidirectional communication network. Through the application of advanced sensing and measurement technology, advanced equipment technology, advanced control methods, and advanced decision support system technology, it achieves the reliable, safe, economical, efficient, friendly environment and safe use of the power grid [1]. The intelligence of the power grid is reflected in six aspects: intelligent substations, power generation, intelligent transmission, intelligent distribution, intelligent electricity consumption, and intelligent scheduling [2]. Among them, the distribution system is a power network system composed of various distribution equipment (or components) and distribution facilities, which converts voltage and directly distributes electrical energy to end users; But the intelligent distribution system is an electrical energy management system with strong professionalism, high degree of automation, ease of use, and high reliability. It is developed according to user needs and follows the standard specifications of the distribution system. Through telemetry and remote control, the rational allocation of loads, optimization of operation, effective energy conservation, and recording of peak and valley electricity consumption have been achieved, providing necessary conditions for energy management [3].

With the continuous development of electronic communication technology and the successive formulation of a series of policies and corresponding technical specifications for energy conservation and emission reduction

[*] State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China. (Corresponding author, YajieLi9@163.com)

[†] State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China. (TaoMing16@126.com)

[‡] State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China. (JiangtaoGuo8@163.com)

[§] State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China. (YuanCao88@126.com)

[¶] State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China. (HongLi287@163.com)

by the country, especially the current development and promotion of technologies and solutions such as cloud platforms and the Internet of Things. Currently, more and more large enterprise users in industries such as construction, electricity, and transportation hope to automatically manage various energy demands and achieve effective energy conservation, alternatively, real-time data on enterprise energy consumption can be obtained through cloud platforms, and distributed monitoring and centralized management of various energy systems can be carried out [4].

2. Literature Review.

The intelligent gateway acts on the network layer of the three-layer architecture of the Internet of Things, controlling sensor devices to collect data downwards and connecting to cloud platforms upwards. Its main functions include data collection, data transmission, and protocol conversion, thereby achieving data interaction between terminal devices and the Internet of Things cloud platform. It is an important component of the Internet of Things system [5]. Feng, C. designed an intelligent industrial IoT gateway that integrates Modbus protocol and MQTT protocol, achieving real-time collection of industrial equipment operation status and data, and sending them to cloud servers through 4G mobile networks via MQTT protocol [6]. Singh, A. conducted research on communication protocols and solved the problem of message interaction and connection between Modbus-485 network and Ethernet [7]. Li, J. designed an Internet of Things gateway based on the actual situation and specific needs of a tea factory. The PLC is responsible for collecting relevant data and sending it to the cloud through the MQTT protocol, achieving the functions of data collection, transmission, and control [8].

Edge computing refers to providing a nearby intelligent service platform on the network side close to the terminal equipment. This platform integrates network, computing, storage, application and other capabilities to meet the digital needs of the Internet of Things industry. The working nature of edge computing is similar to cloud computing. The difference is that cloud computing processes and preprocesses the data collected by gateway devices at the application layer to achieve data persistent storage, while edge computing refers to processing and preprocessing the collected data at the edge of the network. In the implementation process of intelligent gateway, it refers to processing, computing and filtering the collected terminal data in the gateway to improve data transmission efficiency and reduce network resource consumption [9]. Li, X. In view of the network bandwidth problem existing in the greenhouse control system under the traditional cloud computing mode, the edge computing support service platform was built to analyze and process data at the edge side, reducing the pressure of cloud computing and improving the data transmission efficiency [10]. Kumari, P. applied the edge computing framework to the implementation of the Internet of Things gateway. Using the edge computing framework, data can be processed at the edge side, effectively solving the problem of high real-time data processing in solving massive data problems in traditional gateways [11].

By analyzing the structure and functional requirements of power gateway based on edge computing, and building a general service in the gateway through Docker technology, the author proposes a Docker application interaction structure based on message queue telemetry transmission protocol. Through the real-time object detection of the regional network, the cooperation between the edge computing node and the cloud center is realized, the author has designed a power gateway system based on edge computing.

3. Research Methods.

3.1. Edge computing and Requirements.

Edge computing is proposed in the concept of Content Delivery Network, and has obtained preliminary definitions and specifications from the European Telecommunications Association. edge computing organizations have also made industry settings [12]. Through improvement, it is concluded that "edge computing is a computing model that provides computing services at the data source". Edge refers to the resources that devices arrive at the cloud server. In edge computing, on the one hand, sensors send data to the cloud center, and on the other hand, sensors respond to the request of the cloud center. Edge devices are both data producers and users, forming two factions with cloud computing.

Edge computing has the following characteristics:

Localization. The core is to process and store data close to the source. Compared with centralized cloud computing, edge computing can better avoid information leakage caused during data transmission [13].

Timeliness. With the increasing number of intelligent devices, the load of cloud computing is gradually increasing, making it difficult to ensure data timeliness. Deploying resources at the edge of devices for

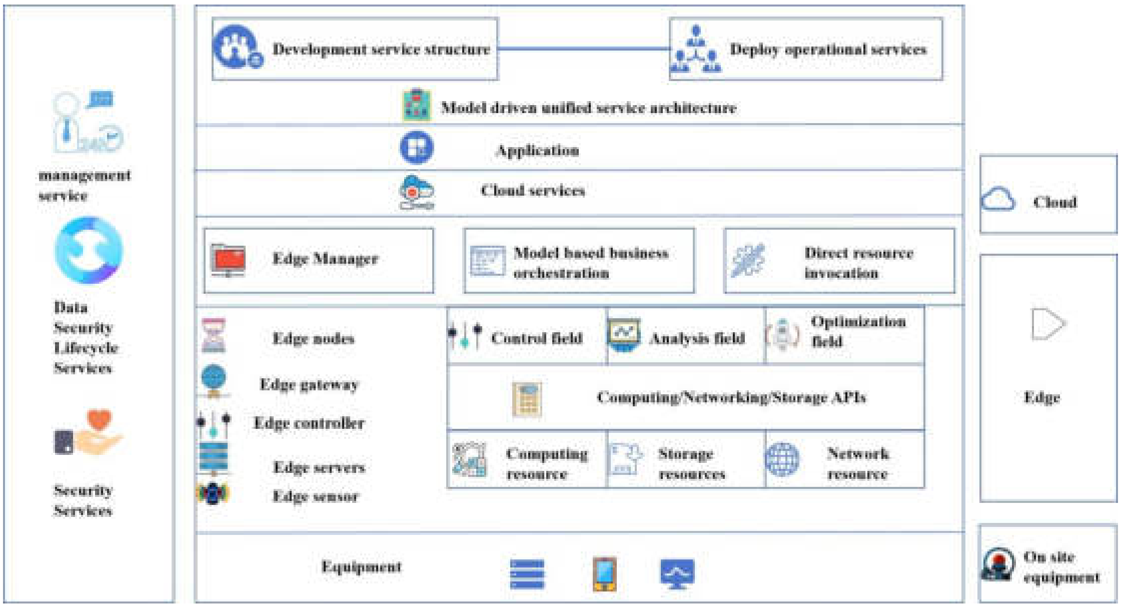


Fig. 3.1: Edge computing 3.0 Structural Framework

Table 3.1: Edge computing equipment requirements

ask	content
Multiple access methods	Due to the complexity of device communication protocols, the gateway is the only entry point for IoT devices, which requires protocol parsing and data transmission. In order to adapt to the data of underlying devices, the gateway needs to be suitable for various protocol interfaces
Cloud connectivity capability	The secure connection between the gateway and the device is a prerequisite for secure data transmission, which is related to the registration, connection method, and push method of the device
Local computing	Edge computing emphasizes processing at the data source, and requires hardware to have certain computing power and storage function

computing can effectively share the burden of cloud centers and improve data timeliness [14].

Low energy consumption. Because the data is complex and centralized, and the power resources consumed by the cloud center exceed 1.5% of the whole society, edge computing can disperse the center data for processing to effectively reduce the center load and energy consumption. Edge computing 3.0 emphasizes the cooperation between the physical world and the digital world to realize the decoupling requirements of software and language, as shown in Figure 3.1.

3.2. Characteristics of IoT gateway equipment and Docker virtual technology. The edge computing node is an important factor in establishing the model. The role of the IOT gateway is to ensure communication and realize network edge management. In edge computing, equipment needs to meet the requirements listed in Table 3.1.

Docker virtualization technology abstracts the network, computing, and storage functions of computers, eliminating hardware limitations on services and enabling users to fully utilize resources. This technology has the advantages listed in Table 3.2.

3.3. Service interaction structure for message queue transmission. The traditional Docker service structure has poor scalability, long resource response time, and complex service invocation methods. The author

Table 3.2: Docker Technical Advantages

advantage	content
Continuous Deployment	During the design process, different testing and development environments can lead to program errors. Virtualization technology ensures that the version configuration of the application is consistent with the requirements, improving development efficiency
Isolation	The system, network, process and other elements of the application are independent of each other, so that when modifying one factor, the other factors will not be affected
safety	Due to its isolation, different programs will not affect each other. Overall, this technology system can only use its own resources, ensuring security

designed a service structure based on Message Queuing Telemetry Transport (MQTT) to improve its orderliness and scalability.

Compared with traditional Docker structures, the structure based on MQTT protocol has the following characteristics:

Scalability: Service interaction relationships are based on topics, and there is no strict sequential logic. When adding or deleting an application, only the corresponding part needs to be modified .

Shared resource release: Two structural interaction modes. When A and B access C simultaneously, in the traditional structure, B needs to wait until the access to A is completed before accessing C, which leads to the inability of the service to respond in real-time.

In traditional structures, if errors occur, multiple applications need to be designed simultaneously, and chain interaction increases the difficulty of eliminating errors; Based on the structure of MQTT, a topic only designs services related to itself, which can effectively delineate boundaries and make it easy to detect errors [15,16].

3.4. Recognition framework based on real-time detection of regional networks. The recognition algorithm adopts the Forward RealTime Object Detection with Region Proposal Networks (Faster RCNN) to improve the calculation speed and result accuracy on the traditional basis. It uses a selection search method to mark the candidate boxes of objects, shorten the generation time of candidate boxes, and improve the calculation speed. The Faster RCNN structure mainly consists of feature extraction, candidate region generation network, investment return rate, and fully connected layers [17].

Feature extraction. Using convolutional neural networks to extract feature maps, the image size relationship of the convolutional layer is shown in equation 3.1, where ($volum_{size}$ represents the size of the convolutional kernel, pad represents the edge zeroing of the image, and stripe represents the step size):

$$Output_{size} = \frac{Input_{size} - volum_{size} + 2 * pad}{stride} + 1 \tag{3.1}$$

The pooling layer uses maximum pooling to reduce the feature map size by 50%, and reduces it to 1/16 of the original size through four pooling layers, the output features of the nth layer are shown in equation 3.2 (β_i^n represents weight, down (x) is matrix sum, b_i^n is bias value, f(x) is Softmax function):

$$x_i^n = f(\beta_i^n down(x_i^{n-1}) + b_i^n) \tag{3.2}$$

Generate a network of candidate regions. The candidate region mechanism of Faster RCNN adopts a generative network, which is different from traditional search methods. The generative network completes end-to-end training through convolutional mode. By annotating candidate boxes through intersection and comparison, objects with low probabilities are removed based on the results. The expression of intersection and union ratio (IoU) is shown in equation 3.3 (A represents the candidate boxes of the generated network, B represents the manually selected switch area):

$$IoU = (A \cap B)/(A \cup B) \tag{3.3}$$

The annotation of candidate boxes based on the intersection and union ratio results is shown in equation 3.4

$$label = \begin{cases} 1 \quad IoU > 0.7 \\ -1 \quad 0.3 \leqslant IoU \leqslant 0.7 \\ 0 \quad IoU < 0.3 \end{cases} \tag{3.4}$$

Candidate box annotation values greater than 0.7 indicate positive samples, while values less than 0.3 indicate negative samples. Candidates between these values will not participate in the experiment.

Based on the results of cls_layer and reg_layer, the optimal solution is selected using non maximum suppression mode, and the loss function is as follows:

$$L(\{pi\}, \{ui\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(pi, pi^*) + \lambda \frac{1}{N_{reg}} \sum_i pi^* L_{reg}(ti \cdot ti^*) \tag{3.5}$$

N_{cls} is the size of the feature map, L_{cls} is the logarithmic loss, L_{reg} is the smoothing loss function, pi is the candidate box switching probability, and pi^* is the label annotation, λ for the loss balance ratio, ti is the offset, and ti^* is the training offset.

3.5. General service settings of data gateway based on edge computing. Implementing a universal service through Docker using Python language, this service framework consists of three layers. The container layer includes file management and logging functions, while the interaction management layer is responsible for exchanging information, communicating with users, and providing universal services.

File management service is the circulation process of general resources in gateway devices. The privacy protection system is responsible for checking privacy protection when users search, store, and transfer files. After confirming security, it will be directed to the search or transfer interface, and privacy monitoring will be carried out throughout the entire process before finally entering the file storage system.

In the file storage process, users need to determine the file storage location or establish a new storage location when storing files. After confirming the file format, the device classifies and stores the files, providing users with the ability to modify the file extension name. Finally, the storage is completed [18].

The retrieval module is responsible for providing users with the function of searching for files. When users need a certain resource, they can search for keywords in the file name or content, and then the computer searches in the file library based on the keywords and pushes them to the user's device. The structural process realizes file upload and download functions. Match the uploaded file with the time through the renaming function, download it through the attachment link, and transfer the file to the user's device.

The file access management process implements file access encryption. Detect users, read user information, and review the allowed access conditions one by one. If one item is not satisfactory, unauthorized users are prohibited from accessing.

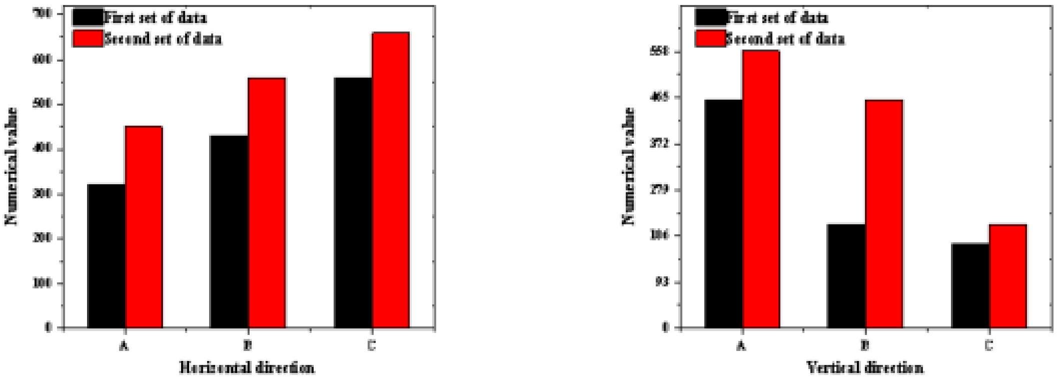
3.6. Coordinates of grounding knife. In order to explore the image processing capability of the gateway, obtain better image samples, and the on/off status of the grounding switch inside the gateway, it is necessary to reflect whether the Faster RCNN model can recognize the position coordinates of the grounding knife. Due to the model of the grounding knife, it is necessary to obtain both vertical and horizontal coordinates, and use equation (6) and Faster RCNN to identify the coordinates of the grounding knife [19].

$$x \cdot cos\theta + y \cdot sin\theta = \rho \to \rho = cos\theta \cdot x + sin\theta \cdot y \tag{3.6}$$

Among them, x and y are the coordinates of the two directions corresponding to the original polar coordinates of the parameters,   and   as an independent variable parameter.

4. Result analysis.

4.1. Grounding knife coordinates and convolutional kernel size. According to the Faster RCNN algorithm based on regional network real-time object detection, when optimizing data network management images, the minimum length data of the monitoring line is obtained by opening and closing the convolutional



(a)

(b)

Fig. 4.1: Parameter Coordinates

kernel size, as shown in Figure 4.1. By analyzing the data, it can be concluded that the above data are all coordinate data of the grounding knife [20].

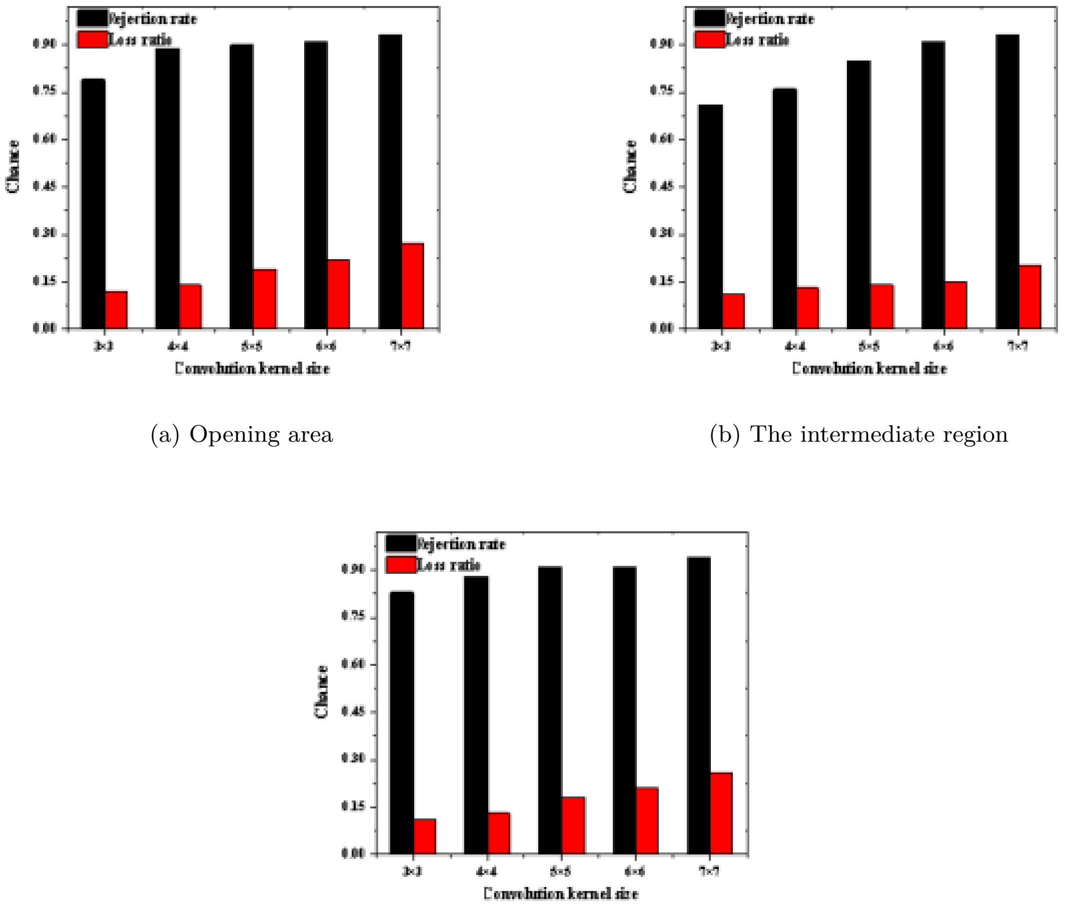
In Figure 4.1(a) (b), area A represents the coordinates of the opening area (320,450) and (460,560), area B represents the coordinates of the middle area (430,560) and (210,460), area C represents the coordinates of the closing area (560,660) and (170,210), and Figures 4.1 (a) and (b) represent the coordinates in two directions. The experimental processing algorithm was adjusted, and the image processing parameters were adjusted to improve the adaptability and accuracy of the algorithm, therefore, it is possible to determine the opening and closing status of the grounding switch inside the gateway based on the results [21].

An excessively large convolution kernel can cause image loss of information and have a negative impact on subsequent algorithms. Divide the range based on parameter coordinates, and change the size of the convolution kernel according to the actual position of the grounding knife to avoid the generation of a single kernel [22]. The author used convolution kernels of different sizes, where the rejection rate represents the proportion of image defects proposed through open close calculations to the total number of defects, and the loss rate represents the proportion of the area lost by the grounding knife contour to the original image, as shown in Figure 4.2(a) (b) (c).

It can be observed that as the size increases, the rejection rate and loss rate of the gateway processing image both increase. The larger the convolution kernel size, the better the image defect removal effect, but at the same time, the edge information loss of the grounding knife also increases [23]. In the opening and closing areas, there is less image information compared to the middle area, so it is more appropriate to use smaller convolution kernels. For the middle area, there are more defects, and comprehensive judgment is needed, compared to 6×6, 5×5 has a lower rejection rate and no significant difference in loss rate; Compared with 6×6, the difference in loss rate between 7×7 and 6×6 is greater than the difference in protrusion rate, so using a convolution kernel size of 6 × 6 is the most suitable [24].

4.2. Grounding knife angle. Due to the different distances between the opening and closing distances of the image processing lens, there is a deviation in the angle of image recognition. Therefore, the three states of the grounding knife are tested. The test data is shown in Figure 4.3.

The above are the angle values for the three states of the grounding knife. Based on the rotation angle of the grounding knife, the region state in which the grounding knife forms three angles can be analyzed. Zone A is in the closed state (92 °, 98 °), zone B is in the middle state (98 °, 175 °), and zone C distinguishes the closed state (175 °, 179 °). The regional network object detection and recognition model can accurately intercept the angle of the grounding knife, and the model based grounding knife recognition model has a certain degree of



(a) Opening area

(b) The intermediate region

(c) Closing area

Fig. 4.2: Image processing results of convolutional kernels with open and closed operations

cyclicity and scalability through transfer learning [25]. According to the test results of the three states of the grounding knife, it can be seen that the recognition algorithm can accurately analyze the state of the grounding knife. Therefore, this algorithm can be used to build a power gateway to ensure information security [26].

5. Conclusion. By learning the basic concepts of edge computing and cloud computing, the author analyzes their advantages and disadvantages, uses a Docker virtual technology as a service deployment tool, and analyzes the adaptability of this technology in edge computing nodes. Through experiments, it was found that Docker has been widely used due to its advantages of lightweight, isolation, and application control. The achievement of building gateway related services has been achieved through the combination of Python language and Docker, and the service interaction structure based on MQTT protocol has solved the complex logic problems existing in traditional Docker structures. The author aims at realizing the operation function of edge computing node in the power gateway, and the research on applying edge computing to the production gateway is slightly insufficient; Moreover, the current application code in Docker is prone to causing user privacy data leakage. Therefore, in future research, how to apply encryption technology to service code and prevent data leakage is the focus of research.

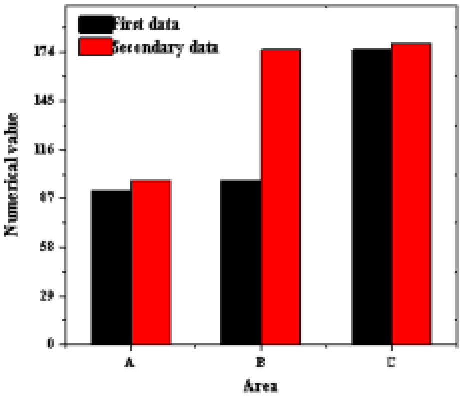


Fig. 4.3: Grounding Knife Status Test

REFERENCES

[1] Liu, D., Liang, H., Zeng, X., Zhang, Q., Zhang, Z., & Li, M. (2022). Edge computing application, architecture, and challenges in ubiquitous power internet of things. Frontiers in Energy Research, 10, 850252.

[2] Minh, Q. N., Nguyen, V. H., Quy, V. K., Ngoc, L. A., Chehri, A., & Jeon, G. (2022). Edge computing for iot-enabled smart grid: The future of energy. Energies, 15(17), 6140.

[3] Qian, Y., Shi, L., Li, J., Zhou, X., Shu, F., & Wang, J. (2020). An edge-computing paradigm for internet of things over power line communication networks. IEEE Network, 34(2), 262-269.

[4] Mehmood, M. Y., Oad, A., Abrar, M., Munir, H. M., Hasan, S. F., Muqeet, H. A. U., & Golilarz, N. A. (2021). Edge computing for IoT-enabled smart grid. Security and communication networks, 2021, 1-16.

[5] Yang, W., Liu, W., Wei, X., Guo, Z., Yang, K., Huang, H., & Qi, L. (2021). EdgeKeeper: a trusted edge computing framework for ubiquitous power Internet of Things. Frontiers of Information Technology & Electronic Engineering, 22(3), 374-399.

[6] Feng, C., Wang, Y., Chen, Q., Ding, Y., Strbac, G., & Kang, C. (2021). Smart grid encounters edge computing: Opportunities and applications. Advances in Applied Energy, 1, 100006.

[7] Singh, A., & Chatterjee, K. (2021). Securing smart healthcare system with edge computing. Computers & Security, 108, 102353.

[8] Li, J., Gu, C., **ang, Y., & Li, F. (2022). Edge-cloud computing systems for smart grid: state-of-the-art, architecture, and applications. Journal of Modern Power Systems and Clean Energy, 10(4), 805-817.

[9] **, W., Xu, R., You, T., Hong, Y. G., & Kim, D. (2020). Secure edge computing management based on independent microservices providers for gateway-centric IoT networks. IEEE access, 8, 187975-187990.

[10] Li, X., Chen, T., Cheng, Q., Ma, S., & Ma, J. (2020). Smart applications in edge computing: Overview on authentication and data security. IEEE Internet of Things Journal, 8(6), 4063-4080.

[11] Kumari, P., Mishra, R., Gupta, H. P., Dutta, T., & Das, S. K. (2021). An energy efficient smart metering system using edge computing in LoRa network. IEEE Transactions on Sustainable Computing, 7(4), 786-798.

[12] Cen, B., Hu, C., Cai, Z., Wu, Z., Zhang, Y., Liu, J., & Su, Z. (2022). A configuration method of computing resources for microservice-based edge computing apparatus in smart distribution transformer area. International Journal of Electrical Power & Energy Systems, 138, 107935.

[13] Khan, L. U., Yaqoob, I., Tran, N. H., Kazmi, S. A., Dang, T. N., & Hong, C. S. (2020). Edge-computing-enabled smart cities: A comprehensive survey. IEEE Internet of Things Journal, 7(10), 10200-10232.

[14] Hamdan, S., Ayyash, M., & Almajali, S. (2020). Edge-computing architectures for internet of things applications: A survey. Sensors, 20(22), 6441.

[15] Al-Dulaimy, A., Sharma, Y., Khan, M. G., & Taheri, J. (2020). Introduction to edge computing. Edge Computing: Models, Technologies and Applications, Institution of Engineering and Technology, London, 3-25.

[16] Yar, H., Imran, A. S., Khan, Z. A., Sajjad, M., & Kastrati, Z. (2021). Towards smart home automation using IoT-enabled edge-computing paradigm. Sensors, 21(14), 4932.

[17] Peruzzi, G., & Pozzebon, A. (2022). Combining lorawan and nb-iot for edge-to-cloud low power connectivity leveraging on fog computing. Applied Sciences, 12(3), 1497.

[18] Jain, S., Gupta, S., Sreelakshmi, K. K., & Rodrigues, J. J. (2022). Fog computing in enabling 5G-driven emerging technologies for development of sustainable smart city infrastructures. Cluster Computing, 25(2), 1111-1154.

[19] Sarker, V. K., Gia, T. N., Ben Dhaou, I., & Westerlund, T. (2020). Smart parking system with dynamic pricing, edge-cloud

computing and lora. Sensors, 20(17), 4669.

[20] Jha, D. N., Alwasel, K., Alshoshan, A., Huang, X., Naha, R. K., Battula, S. K., ... & Ranjan, R. (2020). IoTSim-Edge: a simulation framework for modeling the behavior of Internet of Things and edge computing environments. Software: Practice and Experience, 50(6), 844-867.

[21] Liu, Y., Peng, M., Shou, G., Chen, Y., & Chen, S. (2020). Toward edge intelligence: Multiaccess edge computing for 5G and Internet of Things. IEEE Internet of Things Journal, 7(8), 6722-6747.

[22] Kalyani, Y., & Collier, R. (2021). A systematic survey on the role of cloud, fog, and edge computing combination in smart agriculture. Sensors, 21(17), 5922.

[23] Akhtar, T., & Gupta, B. B. (2021). Analysing smart power grid against different cyber attacks on SCADA system. International Journal of Innovative Computing and Applications, 12(4), 195-205.

[24] Abdulrahman, L. M., Zeebaree, S. R., Kak, S. F., Sadeeq, M. A., Adel, A. Z., 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.

[25] Kumar, A., Alghamdi, S. A., Mehbodniya, A., Webber, J. L., & Shavkatovich, S. N. (2022). Smart power consumption management and alert system using IoT on big data. Sustainable Energy Technologies and Assessments, 53, 102555.

[26] Minoli, D. (2020). Positioning of blockchain mechanisms in IOT-powered smart home systems: A gateway-based approach. Internet of Things, 10, 100147.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 7, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1317–1325, DOI 10.12694/scpe.v26i3.4212

ONLINE EVALUATION OF ERROR STATE OF CURRENT TRANSFORMER BASED ON DATA ANALYSIS

TENGBIN LI[*] QINGCHAN LIU [†] FENGYI ZHENG [‡] YONG CHEN [§] AND ZHAOZHU LI[¶]

Abstract. In order to solve the problem of measurement errors being easily affected by various factors and poor stability of all fiber current transformers, the author proposes an online evaluation of the error status of current transformers based on data analysis. The author proposes a method for evaluating the error status of all fiber current transformers based on correlation analysis: collecting measurement data of three all fiber current transformers at the same measurement point in the converter station, under the constraint of electrical physical correlation, principal component fractal is applied to the measurement data of all fiber current transformers, and error evaluation is mapped to the analysis of changes in Q-statistics. The experimental results indicate that: The method proposed by the author can achieve real-time evaluation of measurement errors in all fiber current transformers, with an evaluation accuracy of up to 0.2 levels. This method greatly improves the evaluation efficiency of measurement errors in all fiber current transformers, reduces the effective power outage time of the power grid, and provides data support for the reliable operation, state prediction, and related technology improvement of all fiber current transformers.

Key words: All fiber current transformer, Measurement error, Standard transformer, Principal Component Analysis

1. Introduction. The online monitoring technology of power equipment is a method of monitoring the insulation status of high-voltage electrical equipment using operating conditions. The important feature is the use of high sensitivity sensors to collect information on the deterioration of electrical equipment insulation during operation, which can accurately monitor the insulation status of operating equipment and provide reliable guarantees for the safe operation of the power system. The secondary output signal of an all fiber current transformer includes the true state information of the primary current and its own measurement error information [1]. Using statistical analysis methods to perform correlation analysis on the secondary output signals of three all fiber current transformers at the same measurement point. Based on the changes in correlation, online evaluation of the error status of all fiber current transformers can be achieved. According to the different analysis objects and application scenarios, correlation analysis methods include binary variable correlation analysis, regression analysis, correlation analysis, and cluster analysis [2].

The ultra-high voltage direct current transmission system has the characteristics of high voltage and large load. The accurate acquisition of primary voltage and current signals is the basis for ensuring the safety, stability, and economic operation of the ultra-high voltage direct current transmission system. The high voltage and high load operation requirements also put forward higher requirements for the reliability and stability of DC measurement equipment. At present, DC measurement equipment mainly includes all fiber current transformers, zero flux current transformers, and DC voltage dividers. Among them, all fiber current transformers have the highest configuration proportion in ultra-high voltage DC transmission systems, exceeding 90%. The current method used for error evaluation and detection of measurement equipment in substations, such as transformers, is the comparison and calibration of standard equipment [3,4]. Because standard equipment has high requirements for operating environment, the method of comparing and calibrating with standard equipment needs to be carried out regularly under power outage conditions in substations. The impact of measurement error changes in all fiber optic current transformers on other equipment in the converter station,

[*] Yunnan Power Grid Co., Ltd. Measurement Center, Kunming, Yunnan, 665000, China (Corresponding author, TengbinLi@126.com)

[†] Yunnan Power Grid Co., Ltd. Measurement Center, Kunming, Yunnan, 665000, China (QingchanLiu7@163.com)

[‡] Yunnan Power Grid Co., Ltd. Measurement Center, Kunming, Yunnan, 665000, China (FengyiZheng8@126.com)

[§] Yunnan Power Grid Co., Ltd. Honghe Bureau, Mengzi, Honghe, Yunnan, 661100, China (YongChen56@163.com)

[¶] Yunnan Power Grid Co., Ltd. Measurement Center, Kunming, Yunnan, 665000, China (ZhaozhuLi9@126.com)

especially energy metering is a long-term cumulative process. However, the essence of regular testing is to conduct a time-domain sampling evaluation of the error status of all fiber current transformers. The testing results have a high degree of randomness. When the measurement error of all fiber current transformers changes during maintenance, this method cannot detect it in a timely manner, and there is a lag in the error detection of all fiber current transformers. On the other hand, all fiber current transformers are mainly used in ultra-high voltage converter stations, with high voltage levels and high primary currents, which also require higher standards. The method of comparing and testing with standards requires a lot of financial and material resources, and has poor economic efficiency. Therefore, it is necessary to conduct online evaluation of the error status of all fiber current transformers without standard equipment, in order to timely detect the deterioration trend of equipment error status [5,6].

2. Literature Review. As a key equipment for current signal sensing in DC transmission systems, accurate measurement of the primary signal is the most critical parameter indicator for measuring the status of all fiber optic current transformers. The accuracy of secondary measurement data from all fiber optic current transformers can be compromised or fail due to performance degradation or failure in any component of the equipment[7]. For instance, Chakraborty, S. S. et al. introduced an enhanced approach that integrates inductance and capacitance within the LVDC bus. This innovative filter diminishes the root mean square (RMS) value of the high-frequency (HF) link current by 29% under conditions of unity power factor operation[8]. Zhao, J. et al. proposed a frequency domain simulation method that does not require circuit synthesis. By combining fast and slow pulse tests, the transmission characteristics of three types of windings in 1000kV GIS CT were cleverly measured. Through disturbance simulation of ultra-high voltage substations, it was found that the radiation loss of the casing has a suppressive effect on high-frequency response, but the influence of frequency related ground impedance can be ignored. Therefore, it is recommended to use a fast damping oscillation wave immunity test not lower than IEC standard level 4 to test the anti-interference performance of secondary equipment in ultra-high voltage substations [9]. Chen et al. presented the harmonic specifications for power system transformers along with the fundamentals of Rogowski coil electronic current transformers. They outlined a straightforward and effective technique for measuring harmonics using Rogowski coil electronic current transformers, and evaluated their performance using high-precision harmonic sources, standard current transformers, and transformer calibrators. The findings demonstrate that this approach offers robust applicability, ease of use, and significantly broadens the spectrum of available harmonic measurement methodologies[10].

In essence, due to the unpredictable nature of power system node conditions within intricate operational settings, existing methodologies fall short in accurately assessing measurement errors across all fiber optic current transformers. To address this challenge, the author suggests an error state assessment approach for all fiber current transformers utilizing correlation analysis. By scrutinizing error patterns in the output signals of all fiber current transformers and leveraging the redundant structural setup characteristic of such transformers in ultra-high voltage converter stations. Data is collected simultaneously from three transformers at identical measurement points within the station. Adhering to electrical-physical correlation constraints, principal component fractal analysis is conducted on the transformer data, with error evaluation mapped to changes analyzed through Q-statistics. Data analysis shows that the method proposed by the author can achieve real-time evaluation of measurement errors in all fiber current transformers, with an evaluation accuracy of up to 0.2 levels. The author's research findings can be extended to the state diagnosis and online evaluation of measurement errors in other DC measurement equipment, and can promote the development of error detection mode in DC measurement equipment from regular passive detection to real-time active self detection [11].

3. Research Methods.

3.1. Error Evaluation Model. Under normal operation, there is a certain deviation between the secondary measurement data of the full fiber current transformer and the true information of the observed primary current signal. This deviation value will be compared and calibrated with standard equipment before the full fiber current transformer is put into operation. It is usually small and stable, and can meet the requirement of 0.2 level measurement accuracy at most. Its mathematical expression can be expressed as follows:

$$x_{ft} = kI_{ft} + v_{ft} + s_{fx} \tag{3.1}$$

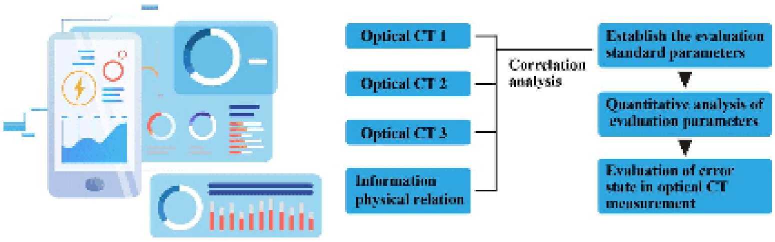

Optical CT 1
Optical CT 2
Optical CT 3
Information physical relation
Correlation analysis
Establish the evaluation standard parameters
Quantitative analysis of evaluation parameters
Evaluation of error state in optical CT measurement

Fig. 3.1: Error evaluation model for all fiber optic current transformers

In equation 3.1, x_{ft} represents the secondary measurement information of the all fiber current transformer; I_{ft} is the observed primary current signal; k is the transmission coefficient of the all fiber current transformer; v_{ft} and s_{fx} are the random errors and systematic errors of all fiber current transformers, respectively. Among them, the random error v_t belongs to a type of free noise, usually satisfying a Gaussian distribution, while the system error s_x is determined by the performance structure of the all fiber current transformer.

$$\begin{cases} x_{1ft} = k_1 I_{ft} + v_{f1t} + s_{f1x} \\ x_{2ft} = k_1 I_{ft} + v_{f2t} + s_{f2x} \\ x_{3ft} = k_3 I_{ft} + v_{f3t} + s_{f3x} \end{cases} \tag{3.2}$$

When three all fiber current transformers operate normally at the same measurement point, the fluctuation of the secondary output signal is mainly due to the normal fluctuation of the next current signal under the influence of the load, that is, the secondary output signal of the three all fiber current transformers is linearly correlated. Therefore, the linear correlation between the secondary output signals of three all fiber current transformers during the initial calibration and operation can be used as the evaluation benchmark. By conducting correlation analysis on the secondary output signals of three all fiber current transformers at the same measurement point, the normal fluctuation of the primary current signal can be separated from the signal deviation caused by the error changes of the all fiber current transformers. By using numerical analysis methods to statistically analyze the error information after peeling, online evaluation of the error status of three sets of all fiber current transformers can be achieved [12]. The error state evaluation model for all fiber current transformers based on correlation analysis is shown in Figure 3.1.

3.2. Basic Principles of Principal Component Analysis. The calculation process of principal component analysis is to collect $X \in \Phi^{n\times 3}$ operating data samples from three all fiber current transformers at the same measurement point, where n is the number of measurement data samples. Decompose the data matrix X

$$X = \hat{X} + E = TP^T + T_e P_e^T \tag{3.3}$$

In equation 3.3, $\hat{X} = TP^T$ represents the principal subspace of the data, which includes the fluctuation information of the primary current under the influence of load; $E = T_e P_e^T$ is the residual subspace of the data, which contains the fluctuation information of measurement errors of all fiber current transformers.

The matrices P and Pe can be obtained by performing singular value decomposition on the covariance matrix R of the running data

$$R = X^T X/(n-1) = [P_1 P_2 P_3] \bigwedge [P_1 P_2 P_3]^T \tag{3.4}$$

In equation 3.4 $\bigwedge = diag(\lambda_1, \lambda_2, \lambda_3)$ and $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$ are the eigenvalues of the covariance matrix R; $[P_1 P_2 P_3]$ is the corresponding feature vector. The larger the eigenvalue, the stronger the correlation between the data.

For the online evaluation requirements of the error status of three all fiber current transformers under the same measurement point studied by the author, the secondary measurement data of all fiber current transformers mainly includes: fluctuation information of primary current under load influence and fluctuation

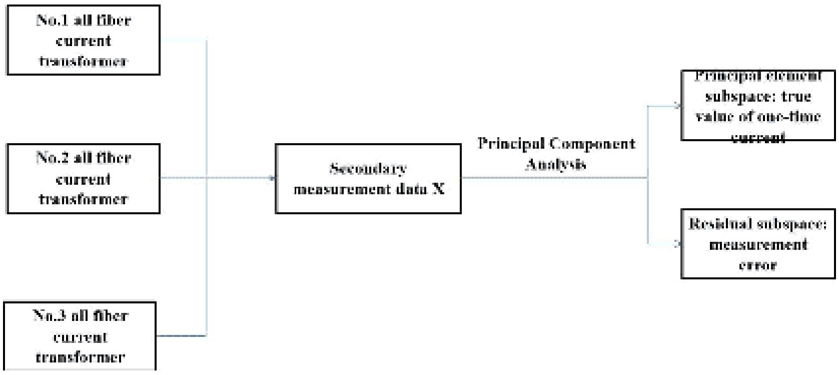

No.1 all fiber current transformer
No.2 all fiber current transformer
No.3 all fiber current transformer
Secondary measurement data X
Principal Component Analysis
Principal element subspace: true value of one-time current
Residual subspace: measurement error

Fig. 3.2: Error detection of all fiber optic current transformers based on principal component analysis

information of measurement error of all fiber current transformers under multiple factors influence. When all components of the fiber optic current transformer are operating normally, the change in measurement error is much smaller than the fluctuation information of the next current affected by the load. By employing principal component analysis (PCA) to break down the measurement data gathered from three all fiber current transformers stationed at identical measurement points, the resulting principal component subspace embodies the genuine essence of the primary current signal. Meanwhile, the residual subspace encapsulates the error-related details stemming from the all fiber current transformers. Figure 3.2 illustrates the error detection process of these transformers facilitated by principal component analysis. For the all fiber current transformer studied by the author, the residual subspace is $P_e = [P_2 P_3]$.

3.3. Calculation of evaluation threshold. According to the above analysis, it can be concluded that using principal component analysis to analyze the measurement data of three sets of all fiber current transformers, the relevant information of measurement errors will be projected into the residual subspace. The degree of deviation in the measurement error of the full fiber current transformer can be determined by calculating the Q-statistic of the operating data in the residual subspace. The calculation method for Q-statistic is

$$Q = (XP_eP_e^T)(XP_eP_e^T)^T = XP_eP_e^TX^T \leqslant Q_c \tag{3.5}$$

In equation 3.5, Q_c is the statistical threshold with a significance level of α. When the Q statistic is greater than this value, it indicates that there is abnormal fluctuation in the measurement error of the all fiber current transformer, which can be calculated according to equation 3.6

$$Q_c = \theta_1 \left[\frac{C_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1} \right]^{\frac{1}{h_0}} \tag{3.6}$$

In equation 3.6: $\theta_i = \sum_{j=\alpha+1}^{3} \lambda_j^i (i = 1, 2, 3)$;$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$; C_α is the critical value of a normal distribution at a detection level of α.

When there is no fluctuation in the measurement error of all fiber optic power transformers, the expected value of the Q-statistic for the measurement data of three sets of all fiber optic current transformers is

$$E(Q) = trE\{I_iP_eP_e^T[I_iP_e]^T\} \tag{3.7}$$

When the error fluctuation of a certain all fiber current transformer is f_i, the mathematical model of the process operation data is

$$X_{fti} = kU_iI_{ti} + U_if_i \tag{3.8}$$

In equation 3.8, U_i is the column vector corresponding to the identity matrix. When the error of an all fiber current transformer at the same measurement point experiences abnormal fluctuations, the expected value of the Q-statistic in the residual subspace of the operating data changes as follows:

$$\Delta E(Q) = E(f_i^2)||P_{ei}||^2 \tag{3.9}$$

According to equation 3.9, when the measurement error of the all fiber current transformer undergoes abnormal changes, the Q-statistic of the measurement data of three all fiber current transformers at the same measurement point will increase, which is positively correlated with the square of the expected error value. Therefore, the state evaluation of measurement errors can be achieved by calculating the Q-statistic of the measurement data of three all fiber current transformers at the same measurement point.

3.4. Anomaly identification methods. When the Q-statistic of the process operation data of an all fiber current transformer exceeds its statistical control threshold Q_α, it indicates that the measurement error of a certain all fiber current transformer has experienced abnormal fluctuations. At this point, it is necessary to further determine which all fiber current transformer has experienced abnormal fluctuations. The contribution plot method is usually used for judgment [13,14].

When the Q-statistic exceeds its threshold, the contribution rate of the measurement data X_o^i of the i-th all fiber current transformer to the Q-statistic is

$$Q_i = e_i^2 = (X_i - \hat{X}_i)^2 \tag{3.10}$$

In equation 3.10, Xi is the column vector corresponding to the measurement data matrix; The column vector corresponding to the main element subspace data matrix of $\hat{X}_i$.

3.5. Error Evaluation Process. Based on the above content, the evaluation process of the error status of three all fiber current transformers at the same measurement point in the converter station using principal component analysis is shown in Figure 3.3. The specific steps are as follows:

1. Collect secondary measurement data of three all fiber current transformers at the same measurement point after calibration and operation, and obtain the process operation data matrix X;
2. Perform singular value decomposition on the data matrix and calculate its eigenvalues $\lambda_1, \lambda_2, \lambda_3$ and its corresponding eigenvectors $[PP_e]$;
3. Selection test confidence α (generally selectable) α= 0.99), calculate the statistical threshold Qc of the Q-statistic according to formula 3.6.
4. Collect secondary measurement data of three all fiber current transformers at the same measurement point during operation, and calculate the Q-statistic of process information according to equation 3.5. If it is less than the statistical threshold Q_c, it indicates that the three all fiber current transformers at the same measurement point to be evaluated are in normal operation; If it is greater than the statistical threshold Q_c, it indicates that there is an abnormal fluctuation in the error state of the all fiber current transformer at this time [15].
5. When the Q-statistic of three all fiber current transformers under the same measurement point to be evaluated exceeds the statistical threshold Q_c, calculate the contribution rates according to equation 3.10, identify the all fiber current transformers with abnormal error fluctuations, and guide relevant personnel in operation, maintenance, and repair work.

3.6. Simulation analysis. Considering that all fiber current transformers are currently mainly used in ultra-high voltage direct current transmission systems with large capacity and high voltage levels, the economic feasibility of conducting error experiments on all fiber current transformers in the laboratory is poor. The author mainly uses numerical simulation methods to analyze and verify the proposed method. At present, the all fiber current transformers in ultra-high voltage converter stations mainly adopt a digital closed-loop technology route, which mainly includes two parts: optical path and detection circuit. The structural schematic diagram of the all fiber current transformer is shown in Figure 3.4. The primary current I(t) forms interference based on the Faraday effect. After photoelectric conversion, signal conditioning, digital to analog conversion, and value

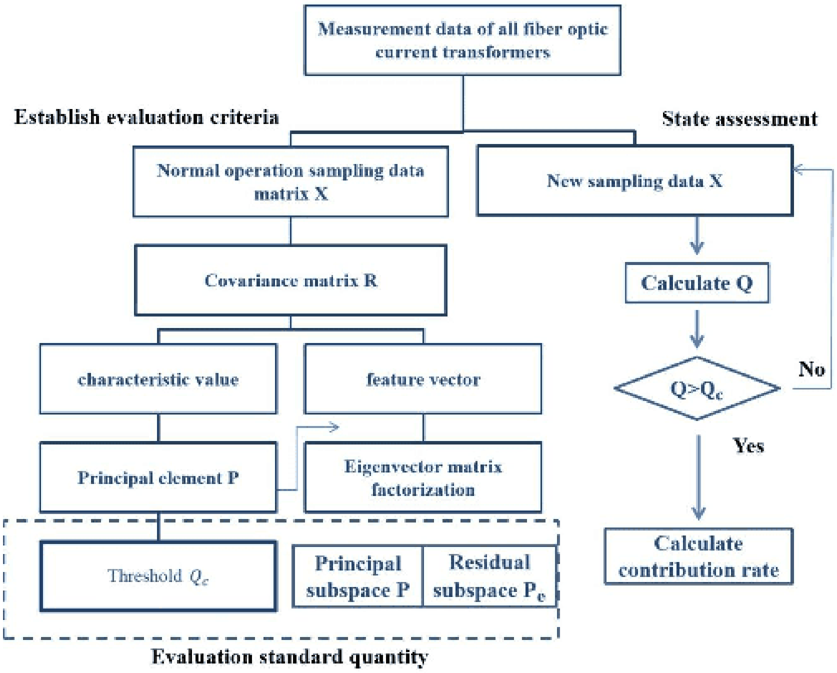

Measurement data of all fiber optic current transformers
Establish evaluation criteria
State assessment
Normal operation sampling data matrix X
New sampling data X
Covariance matrix R
Calculate Q
characteristic value
feature vector
Q>Q_c
No
Principal element P
Eigenvector matrix factorization
Yes
Threshold Q_c
Principal subspace P
Residual subspace P_e
Calculate contribution rate
Evaluation standard quantity

Fig. 3.3: Evaluation process for measurement error of all fiber optic current transformers

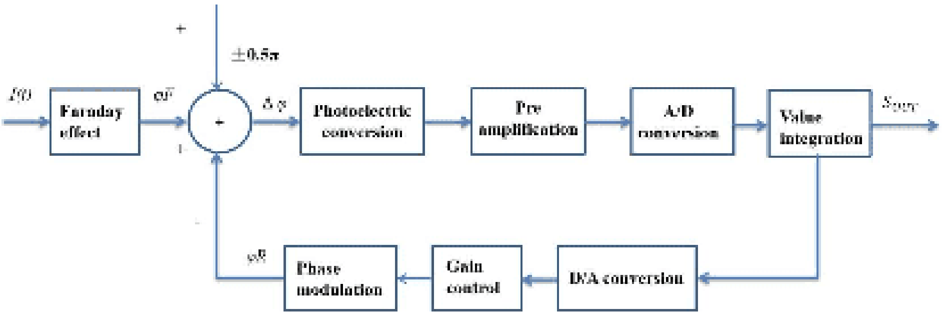


Fig. 3.4: Structural schematic diagram of an all fiber current transformer

integration, a closed-loop feedback signal is formed to compensate for the electro-optic phase feedback, in order to output accurate primary current information [16].

The signal transmission process diagram of the all fiber current transformer is shown in Figure 3.5. Among them, the Faraday magneto optic effect of a single current can generate a phase angle difference process, which can be represented by a proportional coefficient K_f=4NV (N is the number of turns of the sensing fiber, V is the Verdet constant of the sensing fiber). The photoelectric conversion process can be equivalent to a proportional link, with a proportional coefficient of K_1 and a proportional coefficient of K_2 for the preamplifier circuit, the proportional coefficient of the A/D conversion circuit is $K_{AD} = 2^n/U_{ref}$ (n represents the conversion bits of AD, Uref is the reference voltage of the A/D converter), the z-transformation of the value integration is 1/(1-z^{-1}), the proportional coefficient of the D/A converter is K_4, the post gain coefficient is K_5, and the phase modulation is a differential process, its transmission process can be equivalent to $K_m(1 - z^{-1})$. The simplified process flowchart is shown in Figure 3.6.

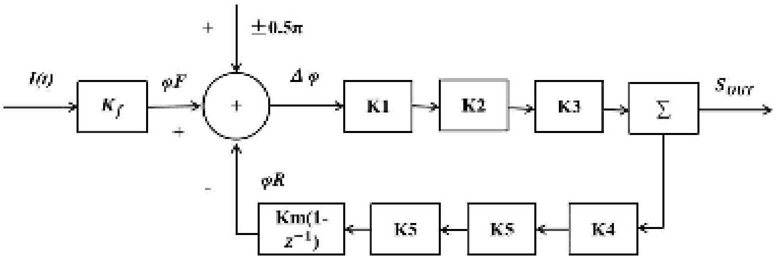


Fig. 3.5: Signal transmission process diagram of full fiber current transformer

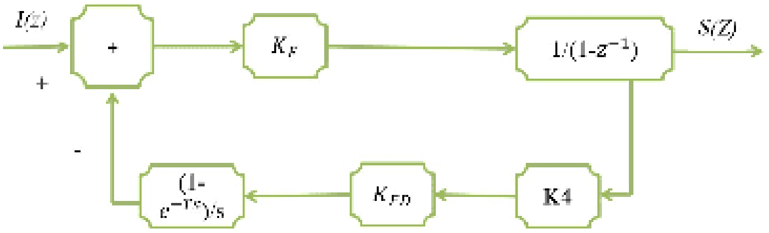


Fig. 3.6: Simplified signal transmission flowchart

In Figure 3.6:

$$K_F = K_f \cdot K_1 \cdot K_2 \cdot K_3 \tag{3.11}$$

The closed-loop transfer function of the all fiber current transformer can be obtained from Figure 3.6 as

$$\frac{S(z)}{I(z)} = \frac{K_F z}{(1 + K_{FD} K_F) z - 1} \tag{3.12}$$

By adjusting the conversion coefficients of various components in the signal transmission system of the all fiber current transformer, various error degradation states of the all fiber current transformer can be simulated [17,18]. The key factor affecting the accuracy of error evaluation for all fiber current transformers under the condition of no standard transformer is the non-stationary time-varying characteristics of the primary current at the node. It is necessary to analyze and verify the method proposed by the author based on the measured data during the operation of all fiber current transformers. The author selected the actual measurement data of a newly calibrated and put into operation all fiber current transformer at a certain converter station (this series of all fiber current transformers has 5 fiber optic sensing turns, 16 A/D conversion bits, and 16 D/A converter bits, resulting in K_F=0.178 and K_{FD}=1.021) for analysis. Simultaneously collect process operation data of three all fiber current transformers at the same measurement point of the converter station. The sampling frequency of this series of all fiber current transformers is 10kHz. The amplitude information of the measurement data is calculated every second, and the average amplitude information of the measurement data during this time period is calculated every 10 minutes. As a sampling point, a total of 7300 sampling points are obtained.

4. Results and Discussion.

4.1. Analysis of Normal Operation Status. According to equation 3.13, calculate three sets of measurement data for simulating all fiber current transformers, and use the first 1000 sets of process operation data as training data to establish an error evaluation model for all fiber current transformers. Use PCA for data analysis, and the model parameters can be obtained as shown in Table 4.1.

Table 4.1: Principal Component Model Parameters

Confidence level	Principal element number	Main component proportion/%	Statistical control threshold	Expected value of statistics
0.99	1	98.8037	0.1402	0.0406

An error state evaluation was conducted on the secondary output data of the three simulated all fiber current transformers mentioned above. The process monitoring of the Q-statistic resulted in abnormal amplitude data accounting for 5.04%, respectively. When considering the Q-statistic and its threshold calculation, the concept of confidence was introduced, and the proportion of abnormal data was randomly distributed. It can be considered that the measurement error of the all fiber current transformer did not fluctuate, which is consistent with the actual situation of the newly calibrated and put into operation of the all fiber current transformer [19].

4.2. Error Analysis. By adjusting the proportion coefficient of the preamplifier circuit of the first all fiber current transformer, a fixed deviation of 0.2% is achieved in its output. Referring to equation 3.13, the measurement data of three sets of simulated all fiber current transformers are calculated accordingly. The error in the secondary measurement data of these transformers is then evaluated, with the process of Q-statistics monitoring. It's observed that the Q-statistic of all three all fiber current transformers surpasses their statistical control limit, indicating an abnormal alteration in their measurement error. Utilizing equation 3.10, the contribution rate of the secondary measurement data from the three transformers to the Q-statistic is computed. Notably, the first all fiber current transformer exhibits the highest contribution to the Q-statistic, suggesting a change in its measurement error, aligning with the real scenario. This method proposed by the author enables accurate evaluation of the measurement error of all fiber current transformers without the reliance on standard devices. The highest evaluation accuracy can meet the evaluation requirements for measurement errors of 0.2 level all fiber current transformers [20].

5. Conclusion. The author proposes an online evaluation of current transformer error status based on data analysis (online monitoring of measuring current transformers). Drawing from an examination of the error traits of all fiber optic current transformers and their deployment specifics within ultra-high voltage converter stations, a novel long-term online monitoring approach for measuring their errors is suggested, grounded in correlation analysis. Initially, the method employs principal component analysis to conduct correlation analysis on the measurement data from three all fiber current transformers stationed at identical measurement points within the converter station. The error fluctuation information of the transformers themselves and the primary current fluctuation information caused by the load are separated, and standard statistics are constructed using normal measurement data for real-time monitoring of all fiber current transformers. Collect actual measurement data of all fiber optic current transformers at the converter station and conduct error simulation analysis. The results show that, this method can quickly evaluate the deterioration status of measurement errors in all fiber current transformers, with a recognition accuracy of up to 0.2 levels. This method greatly improves the evaluation efficiency of measurement errors in all fiber current transformers, reduces the effective power outage time of the power grid, and provides data support for the reliable operation, state prediction, and related technology improvement of all fiber current transformers.

REFERENCES

[1] Wei, B., Xie, Z., Liu, Y., Wen, K., Deng, F., & Zhang, P. (2022). Online monitoring method for insulator self-explosion based on edge computing and deep learning. CSEE Journal of Power and Energy Systems, 8(6), 1684-1696.

[2] Liang, T., Wangwang, L. I., Lei, C., Yongwei, L. I., Zhiqiang, L. I., & Xiong, J. (2022). All-sic fiber-optic sensor based on direct wafer bonding for high temperature pressure sensing. Photon sensor: English version, 12(2), 10.

[3] Rao, Y. (2023). The cornerstone of fiber-optic distributed vibration/acoustic sensing: -otdr. Progress in Optoelectronics, 6(7), 5-8.

[4] Wu, X., Yang, S., Xu, S., Zhang, X., & Ren, Y. (2022). Measurement and correlation of the solubility of sodium acetate in eight pure and binary solvents. Chinese Journal of Chemical Engineering (English Edition), 44(4), 474-484.

[5] Hongrui, L., Xianlong, Z., Zihao, J., Jie, Z., Xiaofei, W., & Ruizhi, Z., et al. (2022). A 16-bit,±10-v input range sar adc with a 5-v supply voltage and mixed-signal nonlinearity calibration. Electronic Journal: English Edition, 31(4), 8.

[6] Li, X. (2023). Novel all-fiber-optic technology for control and multi-color probing of neural circuits in freely-moving animals. Progress in Optoelectronics, 6(7), 1-4.

[7] Liu, H., Deng, Z., Li, X., Guo, L., Huang, D., & Fu, S., et al. (2022). The averaged-value model of a flexible power electronics based substation in hybrid ac/dc distribution systems. CSEE Journal of Power and Energy Systems, 8(2), 452-464.

[8] Chakraborty, S. S., Bhawal, S., & Hatua, K. (2023). Minimization of low frequency current oscillation in resonant link of a solid state transformer by passive filters. IEEE Transactions on Industry Applications, 9(3), 1227-1234.

[9] Zhao, J., Chen, W., Li, K., Teng, Z., Li, N., & Wen, T., et al. (2023). Analysis of conducted disturbance via current transformer due to switch operation of gis disconnector in uhv substation. Journal of Electric Power and Energy Systems, Chinese Society of Electrical Engineering (English), 9(3), 1227-1234.

[10] Chen, L., Chen, H., Zhu, Z., Wu, Y., Zhen, H., & Tong, T., et al. (2023). Study on harmonic metering characteristics of rogowski coil current transformer. AIP Advances, 13(4),14-18.

[11] Peng, Y. J., Lin, C. T., & Chen, Y. M. (2024). Fuzzy data association-towards better uncertainty tracking in clutter environments. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 32(02), 185-207.

[12] Junlei Liu Feng Qian Feng Wang Yun Yang, Feng, Q., Feng, W., & Yun, Y. (2022). Coordinated control method for line steady-state voltage of two-phase source dc transmission system based on closed-loop control. Journal of Power Supply, 20(4), 138-145.

[13] Zhong, Y., Hao, J., Wang, X., Liao, R., Gong, R., & Ding, Y. (2023). Classification method for mechanical defects in gis equipment based on mode function analysis and improved relevance vector machines. CSEE Journal of Power and Energy Systems, 9(2), 790-801.

[14] Cheng, L., He, Z., Liu, J., Yang, Z., Chen, X., & Zhang, Y., et al. (2022). Research on radiated disturbance to secondary cable caused by disconnector switching operation. Energies, 15(5), 1849.

[15] Kang, D. S., Won, W. S., Ahn, B. L., & Lee, B. G. (2022). Fault detection and protection system of electric railway substation. Energy and Power Engineering (David, USA), 16(4), 147-155.

[16] Zhichao Yang Shaozhi Sun Hong Li Mingmin Zhao Peng Zhao Shanshan Lin, Shaozhe, S., Hong, L. I., Mingmin, Z., Peng, Z., & Shanshan, L. (2022). Analysis and measurement of interference current in signal line of electronic transformer signal collector caused by transient operation. Journal of Power Supply, 20(3), 196-203.

[17] Liang, X., Kang, J., & Wang, L. (2022). The relationship between the approximate number system and mathematical abilities: evidence from developmental research. Advances in Psychological Science, 29(5), 827-837.

[18] Dui, H., Zheng, X., Guo, J., & Xiao, H. (2022). Importance measure-based resilience analysis of a wind power generation system:. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 236(3), 395-405.

[19] Ahmad, J., Chang-Hua Lin, Hwa-Dong Liu, Shiue-Der Lu, & Tzu-Hsien Kao. (2022). Analysis of a new tripled boost high voltage gain dc/dc converter with continuous input current. IEEJ Transactions on Electrical and Electronic Engineering, 17(4), 532-538.

[20] Huang, L., Zuyuan, H. E., & Fan, X. (2023). Simplified single-end rayleigh and brillouin hybrid distributed fiber-optic sensing system. Chinese Science: Information Science (English Version), 66(2), 2.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jun 28, 2024
Accepted: Aug 7, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1326–1335, DOI 10.12694/scpe.v26i3.4239

DYNAMIC SCHEDULING OF MULTI-AGENT ELECTROMECHANICAL PRODUCTION LINE BASED ON BIOLOGICAL ITERATIVE ALGORITHM

YAN ZHANG*AND ZIPENG LI†

Abstract. In order to solve the dynamic job scheduling problem in current intelligent machining systems, the author proposes a multi-agent electromechanical production line dynamic scheduling based on iterative algorithms. The author designed a collaborative control method for hybrid micro assembly production lines based on multi-agent technology. Firstly, a mathematical model is used to describe the collaborative control objectives of the production line, and a hybrid micro assembly production line information collection and integration framework is constructed to obtain production line information. By combining the dynamic coordination performance of multi-agent technology, a collaborative control model for production lines is constructed, with the goal of minimizing processing costs as the collaborative control objective. The optimal collaborative control scheme is solved to achieve collaborative control of hybrid micro assembly production lines. The experimental results show that compared with traditional methods, the collaborative control task allocation time obtained by applying this method is shorter, with a minimum value of 15.38 seconds, indicating that this method has higher efficiency in collaborative control of production lines. Compared with traditional methods, the collaborative control task allocation time after applying this method is shorter, effectively reducing the production line processing cost, proving the feasibility of this method.

Key words: Multi Agent technology, Production line, Scheduling, Collaborative control, Mixed Microassembly

1. Introduction. With the development of market economy, manufacturing enterprises are facing a series of challenges. Customers have increasingly high requirements for product quality, and product prices have decreased due to competition. It is urgent for enterprises to reduce production costs and enhance their competitiveness [1]. Production scheduling is an important part of manufacturing systems. Properly handling workshop scheduling problems can save a lot of manpower and material resources for enterprises, enable manufacturing equipment and resources to be more fully utilized, improve production efficiency, and enhance competitiveness. Meanwhile, as the workshop production scheduling problem is a typical NP hard problem, its research has high theoretical and practical significance [2]. With the rise of industrial transformation and technological revolution, intelligent manufacturing has emerged and rapidly developed, becoming one of the main development trends in the current manufacturing industry [3]. The core of intelligent manufacturing is the research and application of intelligent sensing and control equipment, high-end CNC machine tools, and major complete sets of equipment. Smart production lines are one of the main achievements of intelligent manufacturing, which effectively combines intelligent factories with digital workshops and has been widely applied in the manufacturing industry, such as automotive manufacturing, electronic manufacturing, etc. [4]. The production line has advantages such as large-scale production, high efficiency, and low cost, and has become one of the key research directions for the development of national manufacturing industry today.

The development of modern manufacturing industry is very rapid, and the complexity of manufacturing products has also increased, which has put forward higher requirements for production lines [5]. The hybrid micro assembly production line has emerged, which mainly uses hybrid micro assembly technology to connect multiple devices with high density, making the production line develop towards lightweight, miniaturization, high reliability, and low cost trends. The application of hybrid micro assembly technology has reduced the scale of production lines, accelerated work efficiency, and improved intelligence level, providing strong assistance for the development of manufacturing industry [6]. At the same time, the internal structure of the production

*Yellow River Conservancy Technical Institute, Department of mechanical engineering, Henan Kaifeng, 475000, China. (YanZhang5881@163.com)

†Yellow River Conservancy Technical Institute, Department of mechanical engineering, Henan Kaifeng, 475000, China. (Corresponding author, ZipengLi81@126.com)

line has become more complex, posing significant challenges to the control performance of the production line [7]. The concept of Agent originates from the field of artificial intelligence, and its derived Agent technology is an effective way to solve distributed application problems, which has been widely applied in many fields [8]. Agent is a fundamental component of MAS, and the structure and function of Agent itself, as well as the negotiation mechanism and organization method between Agents, have a direct impact on the performance of MAS. Therefore, building an agent with clear functions, reasonable structure, and efficient communication is the fundamental and key step in studying the dynamic scheduling method of flexible manufacturing based on MAS [9].

2. Literature Review.

Under the trend of transforming production organization from a static centralized hierarchical structure to a dynamic distributed network structure, manufacturing systems exhibit new characteristics such as dynamics, complexity, and autonomy. As an application of artificial intelligence, Multi Agent System (MAS) is considered one of the most promising methods for implementing intelligent manufacturing systems and has received widespread attention from scholars both domestically and internationally. In a multi-agent system, agents can autonomously respond to their environment, and each agent is interconnected through distributed and loosely coupled organizational methods, ultimately optimizing the local or global goals of the system through negotiation. The application of multi-agent technology in production scheduling research can enhance the overall scalability and robustness of manufacturing systems, achieve the informatization of production management, and thereby improve production efficiency, safeguarding the transformation and upgrading of the manufacturing industry [10]. Ying, K. C. et al. proposed a mixed integer linear programming (MILP) model and a new metaheuristic algorithm called reinforcement learning iterative greedy (RLIG) algorithm to minimize the completion time of the problem [11]. Duan, W. et al. proposed a constructive greedy heuristic algorithm to efficiently generate and reconstruct solutions to problems. A self-adaptive destruction method has been developed to improve the early development capability of the algorithm and maintain exploration capability in the later stage [12]. Zhou, B. et al. proposed a multi-stage dynamic scheduling algorithm. In the first stage, the static material allocation scheduling problem is decomposed into three sub optimization problems, and dynamic programming algorithms are used to jointly optimize the sub problems to obtain the optimal initial scheduling plan. The superiority of the proposed dynamic scheduling strategy and algorithm was verified through comparative experiments with periodic distribution strategy (PD) and direct insertion method (DI) [13]. Yuan, Z. et al. designed an automatic material scheduling method for industrial production lines based on the FL-NET network. The material scheduling process was divided into two parts: preparation and delivery, and an automatic monitoring platform was established using the FL-NET network structure [14].

The author takes the optimization of Agent electromechanical production line scheduling as the research object. Based on the current difficulties and challenges in the manufacturing industry, an iterative algorithm is used to construct an Agent electromechanical production line scheduling optimization model. The research results aim to provide support for manufacturing enterprises to achieve flexible and intelligent scheduling of production lines, improve production efficiency, adapt to rapidly changing market demands, and promote the manufacturing industry to achieve a more stable position in global competition. The author proposes a dynamic scheduling of multi-agent electromechanical production lines based on iterative algorithms.

3. Method.

3.1. Description of Production Line Scheduling Problems.

For hybrid micro assembly production lines, the collaborative control objective refers to the production line scheduling problem. In order to improve the collaborative control effect of the production line, a mathematical model is used to describe the production line scheduling problem, laying a solid foundation for the subsequent implementation of collaborative control. The scheduling of hybrid micro assembly production lines is essentially a combinatorial optimization problem, which involves finding the optimal collaborative control scheme under multiple constraints to minimize the processing time, minimize processing costs, or maximize processing benefits of the production line. The scheduling of production lines is influenced by various factors, such as funds, manpower, energy, etc. Therefore, the scheduling problem is an NP problem and the ultimate goal of collaborative control.

The author takes the minimum processing cost as the objective function for collaborative control of pro-

Table 3.1: Production Line Information Collection Methods

Acquisition method	Types of information collected	collecting device
Collection card collection	Vibration, sound, acceleration, etc	Data acquisition card
PLC acquisition	I/O status	PPI, Modbus, etc
Ethernet collection	Location information, processing information, auxiliary information, etc	WedService, Modbus, etc

duction lines, and the corresponding mathematical model expression is:

$$\begin{cases} C = min(C_1 + C_2) \\ C_1 = \alpha^* p + \beta^* q + \chi^* t \\ C_2 = \sum_i^n \omega_i |P_i = D_i| \end{cases} \tag{3.1}$$

In equation 3.1, C represents the processing cost of the hybrid micro assembly production line; C_1 represents the cost of product processing; C_2 represents the penalty fee for advance or delay; α^* represents the unit processing cost of the i-th product; p represents the quantity of product processing; β^* represents the unit cost required for replacing the device; q represents the number of replacement components; χ^* represents the cost of moving production line components; t represents the number of movements of the production line components; ω_i represents the penalty coefficient for the i-th product being advanced or delayed; P_i represents the completed quantity of the i-th product; D_i represents the planned processing quantity of the i-th product.

$$\begin{cases} C = min(C_1 + C_2 + C_\delta) \\ C_\delta = \sigma \times d \\ d = \begin{cases} 1 & \text{Unique product processing equipment} \\ 0 & \text{Not in the above situation} \end{cases} \end{cases} \tag{3.2}$$

In equation 3.2, C_2 expressing the penalty function expression; σ represents the penalty coefficient; d represents the auxiliary parameter of the penalty function, with a value of 1 or 0.

The above process uses a mathematical model to describe the production line scheduling problem, which determines the objective function of production line collaborative control, providing a basis for the subsequent collection and integration of production line information.

3.2. Production Line Information Collection and Integration. Based on the above constructed production line collaborative control objective function, the information required for collaborative control of hybrid micro assembly production lines is collected and integrated to prepare for the construction of subsequent collaborative control models.

At present, there are three main methods for collecting information on production lines, as shown in Table 3.1. As shown in Table 3.1, there are differences in the types of information collected by different collection devices applied to different information collection methods. The hybrid micro assembly production line contains a lot of equipment, a complex structure, and a variety of information types. Therefore, this study effectively combines the three information collection methods mentioned above, scientifically and reasonably allocates production line information collection tasks, in order to achieve the best production line information collection effect [15]. Applying Microsoft to build a hybrid micro assembly production line information collection framework, connecting heterogeneous devices, breaking the phenomenon of ”digital islands”, achieving information transmission between heterogeneous devices, providing convenience for collaborative control of production lines, and also assisting in information collection and transmission. The information collection framework for hybrid micro assembly production lines is shown in Figure 3.1.

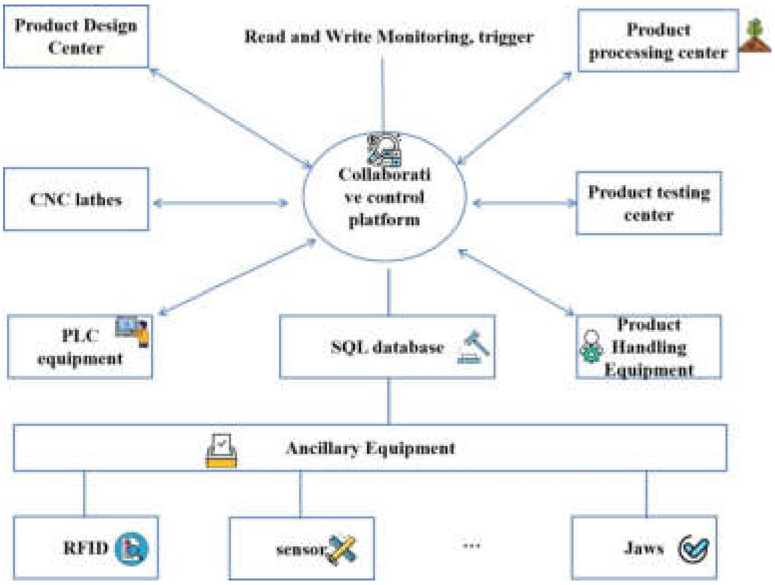

Product Design Center
Read and Write Monitoring, trigger
Product processing center
CNC lathes
Collaborative control platform
Product testing center
PLC equipment
SQL database
Product Handling Equipment
Ancillary Equipment
RFID
sensor
...
Jaws

Fig. 3.1: Schematic diagram of production line information collection framework

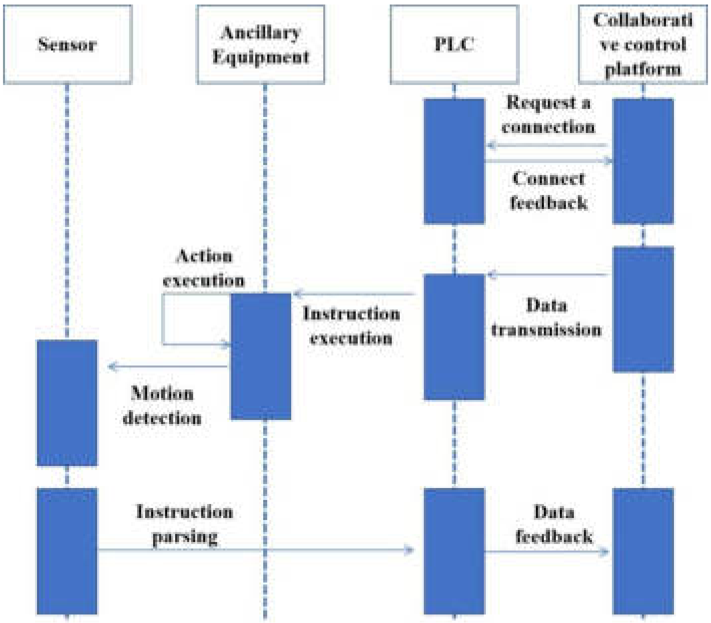


Fig. 3.2: Schematic diagram of information exchange logic for collection devices

As shown in Figure 3.1, in the framework construction, there are many production line information collection devices. In order to ensure the integrity of the collected information, it is necessary to set up the interaction logic of the collection device information, as shown in Figure 3.2.

Interact and integrate the collected information according to the logic shown in Figure 2, and store the integrated production line information in an SQL database, providing convenience for subsequent data applications.

3.3. Collaborative Control Model Construction. Based on the collected and integrated production line information mentioned above, a collaborative control model is constructed using multi-agent technology to provide support for the final implementation of collaborative control.

In multi-agent technology, there are multiple agents with different objectives that can be dynamically coordinated to solve corresponding problems. Multi Agent technology has strong flexibility, adaptability, and reliability, occupying a crucial position in the field of control, and is also one of the main means to solve NP problems today.

The dynamic coordination steps of multi-agent technology are as follows:

Step 1: Release collaborative control task information for the production line. It should be noted that each individual Agent can publish production line collaborative control task information, and a single Agent can publish multiple task information;

Step 2: Agent dynamically coordinates team building. Based on factors such as the collaborative control task information of the production line and the functions of individual agents, select suitable agent individuals to participate in the dynamic coordination team of agents, and construct the team through established rules until the collaborative control task of the production line is completed;

Step 3: Case based reasoning. Based on the published collaborative control task information, search for similar successful cases in the historical records. If there are similar successful cases, carry out dynamic coordination operations according to the cases; If there are no similar successful cases, further accurate selection of coordination attitude is needed;

Step 4: Coordinate attitude selection. This study uses fuzzy number $\xi_\theta(t)$ to describe the coordination attitude of individual agents towards publishing collaborative control task information. The expression for fuzzy number $\xi_\theta(t)$ is:

$$\xi_\theta(t) = \frac{L_0(\theta)}{L_0(\xi)} \tag{3.3}$$

In equation 3.3, $L_0(\theta)$ and $L_0(\xi)$ respectively represent the profit values of Agent individuals θ and ξ in the dynamic coordination process.

According to equation 3.3, select the coordination attitude based on the calculation results, and the specific rules are as follows:

$$\begin{cases} \xi_\theta(t) = 0 \quad \text{Completely selfish} \\ 0 < \xi_\theta(t) < 1 \quad \text{Partial collaboration} \\ \xi_\theta(t) = 1 \quad \text{Complete collaboration} \\ \xi_\theta(t) > 1 \quad \text{Compromise} \end{cases} \tag{3.4}$$

Step 5: Select appropriate coordination strategies based on the coordination attitude determined in Step 4. When $\xi_\theta(t)$=0, the agent's individual coordination attitude is completely selfish, which is called competitive coordination, and the optimal coordination strategy is game theory; When $\xi_\theta(t) \neq 0$, the individual coordination attitude of the agent is collaborative, which is called collaborative coordination. The optimal coordination strategy is partial global planning or FA/C method;

Step 6: Individual Agent Learning. In this step, individual agents should learn coordination strategies, coordination content, and coordination tasks for similar successful cases, in order to complete the corresponding production line collaborative control tasks;

Step 7: Guide the individual agents who have completed the learning to carry out collaborative control tasks on the production line, and disband the dynamic coordination team of the agents when the collaborative control tasks are completed [16].

Through the above process, it can be seen that multi-agent technology has strong dynamic coordination ability, which can provide great support for collaborative control of hybrid micro assembly production lines. Collaborative control of production lines is a complex and cumbersome problem. The author applies the dynamic coordination ability contained in multi-agent technology to construct a collaborative control model for production lines, as shown in Figure 3.3.

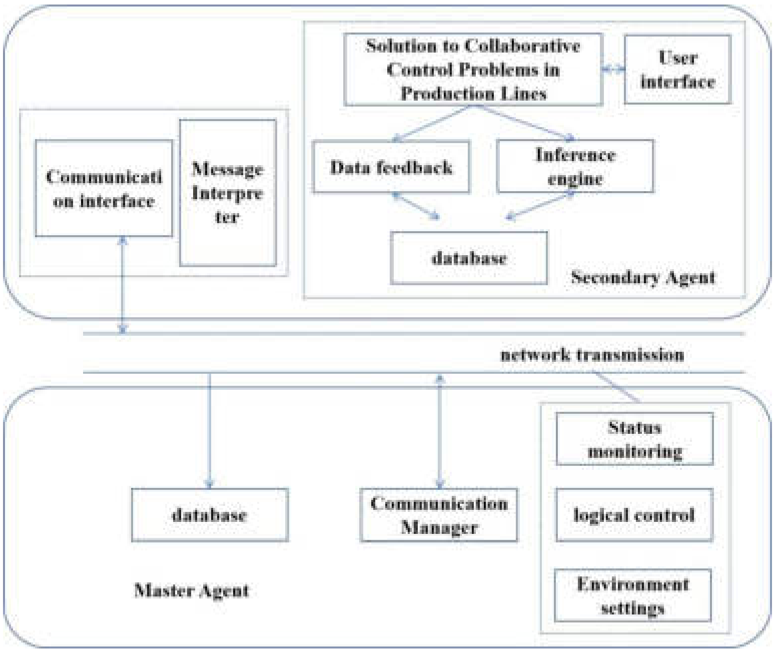


Fig. 3.3: Schematic diagram of collaborative control model for production line

Table 3.2: Collaborative Control Plan Table

Collaborative control steps	Collaborative control conditions	Collaborative control scheme
Step 1	$t_1 > t_2$	$F_1 > F_2$
Step 2	$t_1 < t_2$	$F_1 < F_2$
Step 3	$t_1^* < t_2^*$and $t_1 = t_2$	$F_1 > F_2$
Step 4	$t_1^* > t_2^*$and $t_1 = t_2$	$F_1 < F_2$
Step 5	$t_1^* = t_2^*$and $t_1 = t_2$	$F_1 = F_2$

In addition, a mathematical model is used to represent the internal structure of the Agent:

$$Agent = \{A, MK, D, I, S, R\} \tag{3.5}$$

In equation 3.5, A, M, K, D, I, S, R respectively represent the properties, solutions, relevant regulations, relevant data, reasoning process, information transmission rules, and information reception rules of the Agent.

The above process completed the construction of the collaborative control model for the production line and demonstrated the dynamic coordination performance of multi-agent technology, providing model support for the subsequent implementation of collaborative control.

3.4. Implementation of Collaborative Control on Production Lines. Based on the collaborative control model of the production line constructed above, with the lowest processing cost as the collaborative control objective, the optimal collaborative control scheme is solved to achieve collaborative control of the hybrid micro assembly production line [17].

Applying multi-agent technology to solve the objective function of collaborative control in production lines, the collaborative control scheme is obtained as shown in Table 3.2.

In Table 3.2, t1 and t_2 represent the standard processing costs of the two products; t_1^* and t_2^* represent the processing costs of products with different priority levels; F_1 and F_2 represent the priority of collaborative

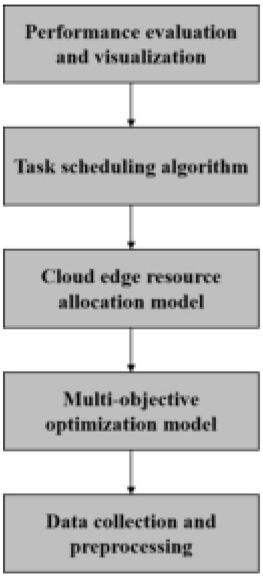

Performance evaluation and visualization
Task scheduling algorithm
Cloud edge resource allocation model
Multi-objective optimization model
Data collection and preprocessing

Fig. 3.4: Technical framework diagram of scheduling method

control.

In summary, the author has implemented collaborative control of hybrid micro assembly production lines based on multi-agent technology, minimizing production line processing costs and contributing to the sustainable development of the manufacturing industry [18].

3.5. multi-objective task scheduling for production lines. The author's dynamic scheduling technology framework for multi-agent electromechanical production lines based on iterative algorithms can be divided into several key components. The technical framework diagram of the scheduling method is shown in Figure 3.4.

In the Data Collection and Preprocessing stage, collect data from IoT devices, sensors, and manufacturing resources to provide necessary information for task scheduling. The data may include processing time, machine availability, energy consumption rates, and any other decision related information. Preprocessing may involve cleaning up data, handling missing values, and normalizing or converting data into a format suitable for further processing.

3.6. Collaborative control performance testing of production lines. In order to verify the application performance of the collaborative control method for hybrid micro assembly production lines based on multi-agent technology, the traditional data-driven real-time monitoring and optimization control technology for production lines is used as a comparative method, and the following comparative experiments are designed [19]. The experimental platform is the foundation for testing the collaborative control performance of hybrid micro assembly production lines. According to the performance testing requirements, an experimental platform was built using conveyor belts, loading boxes, sensors, industrial cameras, etc., as shown in Figure 3.5.

As shown in Figure 3.5, the experimental platform includes two conveyor belts and is equipped with industrial cameras and various sensors to obtain production line related information and prepare for collaborative control of the production line. In addition, industrial cameras need to be combined with photoelectric sensors for application. Once the photoelectric sensor senses product information, it transmits the signal in real time to the industrial camera, triggering it to collect information. Experimental equipment is also one of the key influencing factors to ensure the smooth progress of experiments. The main equipment for designing the experiment is a collaborative control server and communication equipment. Among them, the collaborative control server undertakes the tasks of data processing and forwarding, while the communication equipment undertakes the tasks provided by the high-quality communication environment. The experiment uses Ethernet switches

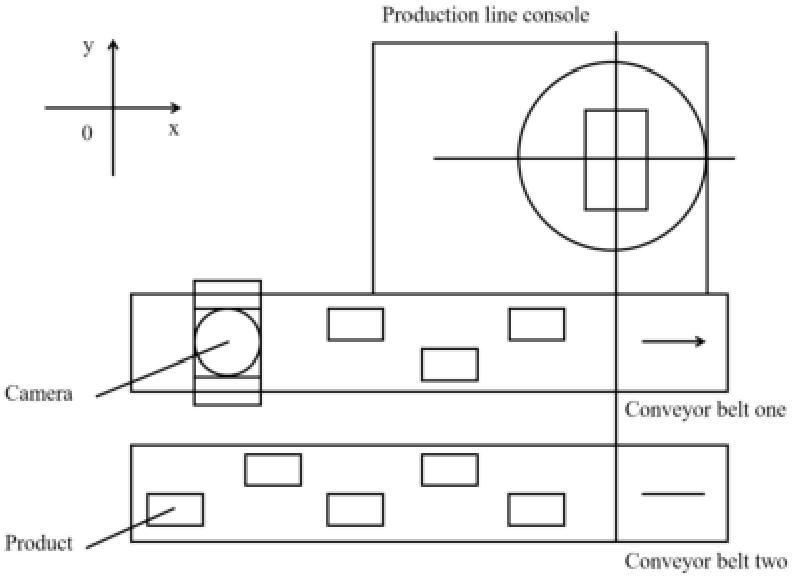


Fig. 3.5: Schematic diagram of experimental platform

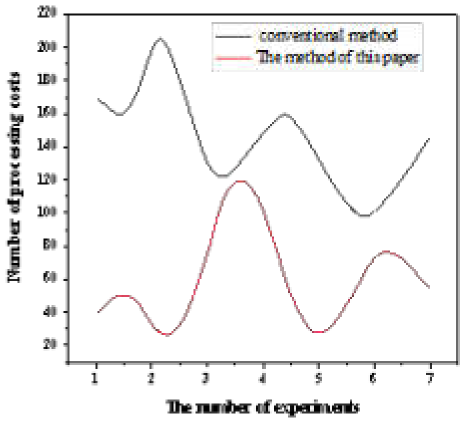


Fig. 4.1: Production Line Processing Cost Data Chart

as communication devices, which can provide a stable and high-quality communication environment within a certain range, providing strong support for performance testing [20].

4. Results and Discussion. Based on the experimental platform built above and the selected experimental equipment, conduct collaborative control performance testing on the hybrid micro assembly production line. In order to quantify the application performance of the proposed method, collaborative control task allocation time and production line processing cost were selected as evaluation indicators. The specific experimental results analysis process is shown in Figure 4.1.

The allocation time of collaborative control tasks indirectly reflects the efficiency of collaborative control in production lines. In general, the shorter the allocation time of collaborative control tasks, the higher the efficiency of collaborative control; On the contrary, the longer the allocation time of collaborative control tasks,

Table 4.1: Collaborative Control Task Allocation Time Data Table

Number of experiments	Author's method	traditional method
1	20.34s	26.53s
2	15.38s	28.41s
3	18.32s	29.23s
4	17.45s	30.16s
5	18.60s	35.05s
6	16.23s	28.53s
7	18.73s	35.27s

the lower the efficiency of collaborative control.

The collaborative control task allocation time data was obtained through experiments, as shown in Table 4.1. As shown in Table 4.1, compared with traditional methods, the collaborative control task allocation time obtained by applying this method is shorter, with a minimum value of 15.38 seconds, indicating that this method has higher efficiency in collaborative control of production lines. The processing cost of the production line directly reflects the effectiveness of collaborative control on the production line. In general, the lower the processing cost of a production line, the better the collaborative control effect; On the contrary, the higher the processing cost of the production line, the poorer the collaborative control effect. The production line processing cost obtained through experiments is shown in Figure 4.1. As shown in Figure 4.1, compared with traditional methods, the production line processing cost obtained by applying this method is smaller, with a minimum value of 240000 yuan, indicating that the collaborative control effect of this method on the production line is better. The above experimental results show that compared to the comparative methods, the collaborative control task allocation time obtained by applying this method is shorter, and the production line processing cost is lower, fully confirming that this method has better collaborative control performance.

5. Conclusion. The author proposes a dynamic scheduling method for multi-agent electromechanical production lines based on iterative algorithms. This study applies multi-agent technology to propose a new collaborative control method for hybrid micro assembly production lines, which greatly shortens the allocation time of collaborative control tasks, reduces production line processing costs, provides more effective method support for collaborative control of production lines, and also provides new ideas and theoretical references for collaborative control research.

Acknowledgements. Science and Technology Research Project of Henan Provincial Department of Education: Research on Key Technologies for Local Continuous Plastic Forming of Bolt Balls Number 202102210265.

REFERENCES

[1] Li, W., Han, D., Gao, L., Li, X., & Li, Y. (2022). Integrated production and transportation scheduling method in hybrid flow shop. Chinese Journal of Mechanical Engineering, 35(1), 1-20.

[2] Liu, Q., Gao, Z., Li, J., Li, S., & Zhu, L. (2023). Research on optimization of dual-resource batch scheduling in flexible job shop. Computers, Materials, and Continuum (in English), 76(8), 2503-2530.

[3] Nanfeng, M. A., Yao, X. F., & Wang, K. S. (2022). Current status and prospect of future internet-oriented wisdom manufacturing. SCIENTIA SINICA Technologica, 52(1), 55-75.

[4] Zhao, L., Shao, J., Yuqi, Q. I., Chu, J., & Feng, Y. (2023). A novel model for assessing the degree of intelligent manufacturing readiness in the process industry: process-industry intelligent manufacturing readiness index (pimri). Frontiers of Information and Electronic Engineering: English Version, 24(3), 417-432.

[5] Hewamanne, S. (2023). Invisible bondage: mobility and compulsion within sri lanka's global assembly line production. Ethnography, 24(1), 85-105.

[6] Alam, S., Dhamija, P., & Ziderman, A. (2022). Human resource development 4.0 (hrd 4.0) in the apparel industry of bangladesh: a theoretical framework and future research directions. International Journal of Manpower, 43(2), 263-285.

[7] Sun, B., Dai, J., Huang, K., Yang, C., & Gui, W. (2022). Smart manufacturing of nonferrous metallurgical processes: review and perspectives. International Journal of Minerals, Metallurgy and Materials, 29(4), 611-625.

[8] Mubarak, M. A. (2023). Sustainably developing in a digital world: harnessing artificial intelligence to meet the imperatives

of work-based learning in industry 5.0. Development and Learning in Organizations: An International Journal, 37(3), 18-20.

[9] Xu, Z., Liu, X., Cao, J., & Song, M. (2022). Fixed-time bipartite consensus of nonlinear multi-agent systems under directed signed graphs with disturbances. Journal of the Franklin Institute, 359(6), 2693-2709.

[10] Shivdas, R., & Sapkal, S. (2023). Proposed composite similarity metric method for part family formation in reconfigurable manufacturing system. The International Journal of Advanced Manufacturing Technology, 125(5-6), 2535-2548.

[11] Ying, K. C., & Lin, S. W. (2023). Reinforcement learning iterated greedy algorithm for distributed assembly permutation flowshop scheduling problems. Journal of ambient intelligence and humanized computing, 19(2), 1367-1394.

[12] Duan, W., Kang, Q., Kang, Y., Chen, J.,& Qin, Q. (2022). A simple and effective iterated greedy algorithm for structural balance in signed networks. International Journal of Modern Physics, B. Condensed Matter Physics, Statistical Physics, Applied Physics, 19(8), 138-148.

[13] Zhou, B., & Wen, M. (2023). A dynamic material distribution scheduling of automotive assembly?line considering material-handling errors. Engineering Computations, 40(5), 1101-1127.

[14] Yuan, Z., & Zhu, X. (2022). Material scheduling method of automated industrial production line based on fl-net network. International Journal of Internet Manufacturing and Services(3), 8.

[15] Weng, T., Xie, Y., Chen, G., Han, Q., Tian, Y., & Feng, L., et al. (2022). Load frequency control under false data inject attacks based on multi-agent system method in multi-area power systems:. International Journal of Distributed Sensor Networks, 18(4), 4610-4618.

[16] Zhou, B., & Wen, M. (2023). A dynamic material distribution scheduling of automotive assembly?line considering material-handling errors. Engineering Computations, 40(5), 1101-1127.

[17] Wang, X., Zhang, G., Li, Y., & Qu, N. (2023). A heuristically accelerated reinforcement learning method for maintenance policy of an assembly line. Journal of Industrial and Management Optimization, 19(4), 2381-2395.

[18] Yan, X., Zuo, H., Hu, C., Gong, W., & Sheng, V. S. (2023). Load optimization scheduling of chip mounter based on hybrid adaptive optimization. Modeling and Simulation of Complex Systems, 3(1), 11.

[19] Hewamanne, S. (2023). Invisible bondage: mobility and compulsion within sri lanka's global assembly line production. Ethnography, 24(1), 85-105.

[20] Leng, J., Sha, W., Lin, Z., Jing, J., Liu, Q., & Chen, X. (2023). Blockchained smart contract pyramid-driven multi-agent autonomous process control for resilient individualised manufacturing towards industry 5.0. International Journal of Production Research, 61(13), 4302-4321.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jul 3, 2024
Accepted: Aug 7, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1336–1347, DOI 10.12694/scpe.v26i3.4311

RESEARCH ON IDENTIFICATION AND DETECTION OF UNSAFE BEHAVIORS OF CONSTRUCTION WORKERS BASED ON DEEP LEARNING

MEIYU ZHANG[*], HONGMING CHEN[†], AND XUEFENG HAN[‡]

Abstract. In order to improve the safety management level of construction sites, prevent and reduce the occurrence of building safety accidents, this article uses deep learning methods to study these unsafe behavior recognition and detection techniques. The most typical hazardous behavior is not wearing a safety helmet. However, on-site personnel often neglect to wear helmets due to various reasons. In this study, the target detection algorithm is applied to monitor helmet-wearing. The YOLOX algorithm is selected as the basic detection model and improved by combining the construction site environment and helmet detection characteristics, meeting the real-time monitoring needs of helmet-wearing. Comparison experiments before and after improvement were conducted on the self-constructed helmet dataset, verifying the performance of the improved YOLOX network model. The results show that the average accuracy of the enhanced network model on the helmet-wearing dataset increased to 89.12%, showing a better detection effect.

Key words: deep learning, YOLOX algorithm, sensory field, unsafe behaviour, target detection, building construction, behavior recognition

1. Introduction. Currently, the safety situation in building construction in China remains a serious concern, as the frequency of safety accidents continues to be high, resulting in elevated numbers of incidents and fatalities. [1]. The majority of these accidents are a result of unsafe behaviors exhibited by construction workers. These workers often engage in long hours of high-intensity physical labor, leading to fatigue and subsequently lazy and risky behaviors. Studies have indicated that traumatic brain injuries resulting from falling objects accounted for 24% of total construction worker accidents, with most of these fatal accidents attributed to a lack of safety helmet use [2]. Previous studies have explored the influencing factors, mechanisms, and pre-control methods of construction workers' unsafe behaviors, but have lacked effective strategies for directly controlling and correcting these behaviors [3]. Meanwhile, many scholars have conducted research on helmet-wearing state detection algorithms based on deep learning methods. In 2018, Fang et al [4] tried for the first time to apply the Faster R-CNN algorithm to helmet-wearing detection, and although the algorithm made some progress in improving detection accuracy, it could not meet the real-time requirements. In 2020, Liang et al [5] based on the YOLOv3 algorithm for helmet detection, due to its relatively single dataset, resulting in poor generalization of the model; in 2023, Qi Zezheng et al [6] used a hybrid pooling optimization spatial pyramid pooling (SPP) module (SPP) in the form of tandem pooling in YOLOv5s and embedded a coordinate attention mechanism in the slicing module, although the detection accuracy of the model is improved, the model is more complex and unfavorable for deployment. This paper aims to address the limitations of the aforementioned research by utilizing a deep learning platform to construct the YOLOX (You Only Look Once X) target detection network model. Additionally, the characteristics of the construction site environment are taken into consideration, and a structural reparameterization model is introduced to enhance the overall network's feature expression capability and computing speed. Furthermore, a sliding window transformation network is incorporated to improve the network's field of perception. The decoupled detector head is also further decoupled to enhance the model's feature extraction ability. The research focuses on developing a method for detecting helmet-wearing among construction personnel, to enable pre-warning, normal detection, and standardized management for construction safety. Two-Stages target detection algorithms based on candidate regions and One-Stage target detection algorithms based on regression are two types of mainstream target detection methods at this stage.

[*] Nanjing Tech University, China
[†] Nanjing Tech University, China
[‡] Nanjing Tech University, China

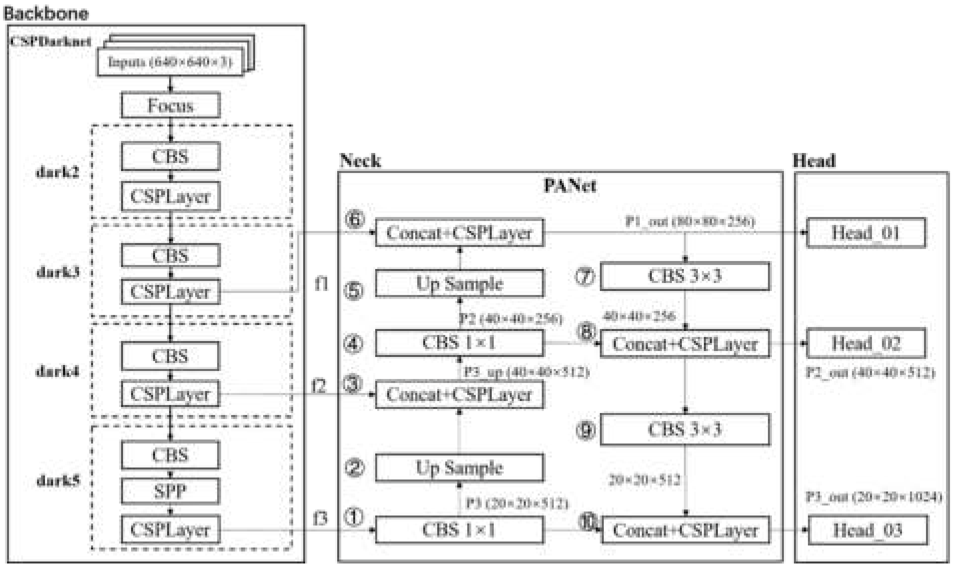


Fig. 2.1: Network structure of YOLOX.

Two-Stages algorithms mainly include R-FCN [7], Faster-RCNN [8], Mask-RCNN [9], and other detection algorithms. These algorithms need to generate candidate regions first, and then classify and localize the targets in the candidate regions; one-stage algorithms mainly include the YOLO (You Only Look Once) series, SSD [10], and other algorithms, which don't generate candidate frames but directly transform the localization problem of the candidate frames into a regression problem to be dealt with, and the whole detection process makes use of the end-to-end (end-to-end) detection of objects. to-end) direct regression of the object's category and location [11]. Compared with the two-stage algorithm, the single-stage algorithm has a faster processing speed and is suitable for real-time application scenarios [12].

2. YOLOX Network Structure. The network structure of the YOLOX algorithm model has three main parts, Backbone, Neck, and Head (as shown in Figure 2.1)[13].The Backbone network is the foundation of its entire architecture, effectively extracting features from input images through a combination of convolutional and pooling layers. These features are crucial for subsequent object detection as they provide basic information and contextual relationships of the image. The Neck section integrates feature information from different dimensions together, and the different focuses of the three outputs are more conducive to detecting targets at different scales. The Head network of YOLOX is the core of the entire object detection model, responsible for generating bounding boxes, category probabilities, and object confidence scores. These prediction results are the final judgment of the model on the target position and category in the input image.

3. Network Modelling Improvement. At this stage, the mainstream target detection methods can be divided into two types: two-stage (Two-Stages) target detection algorithms based on the candidate region and one-stage (One-Stage) target detection algorithms based on regression. From the results of a large number of studies, although the accuracy of the One-Stage target detection algorithm is slightly lower than that of the Two-Stage algorithm under the same circumstances, the detection speed has increased, which can better meet the timeliness of construction site inspection. The task of target detection involves identifying and localizing object information within an image. In the realm of image processing, the focus has shifted towards utilizing deep learning methods over traditional image processing techniques for target detection. When detecting workers wearing safety helmets in a complex construction environment, it is essential to consider the intricacies of the building construction site.

The analysis of site environment characteristics and safety helmet detection is outlined below:

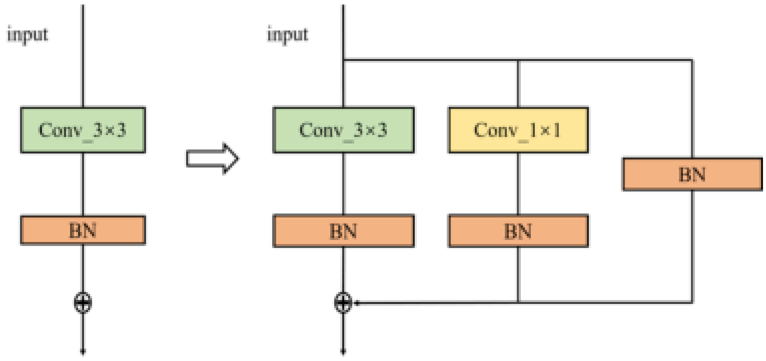


Fig. 4.1: Multi-branch structure of RepVGG.

1. The building construction site has unstable light and cluttered background, and the workers' location distance and occlusion are all uncertain, and these factors will affect the effectiveness of target detection.
2. The helmet target is small, so the detection accuracy of the deep learning target detection algorithm is required to be high. The current deep learning algorithm still has some difficulties in dealing with small targets, high-density targets, and other scenes.

Based on the analysis above, this paper modified the Backbone, Neck, and Head structure of YOLOX to enhance the model's feature extraction capability for detecting workers wearing safety helmets in building construction environments. This modification aims to improve the detection effect and reliability of the model.

4. Introduction of structural reparametric modeling. RepVGG (Re-parameterization VGG) [14], i.e., using the idea of structural re-parameterisation, uses a multi-branch structure to increase the number of parameters that can be computed during network training to improve performance and converts it to a single-path structure during network inference thereby increasing the speed of computation and reducing the memory. RepVGG uses the original VGG network structure as the backbone and makes structural improvements.

Figure 4.1 shows the transformation of the RepVGG-based VGG network into a multi-branch parallel structure during the training phase. It can be seen that the original single-path structure on the left mainly contains Conv_3×3 and BN layers, and the Conv_1×1 residual branch and identity residual branch are introduced on the right. The simple residual structure is transformed into a complex residual structure with the addition of multiple branches, and the network is transformed from a single flow path into multiple flow paths. Training such a network is equivalent to training multiple networks, and thus the parameters that can be computed by the model are greatly increased. This not only enhances the representation ability in the deep network of the model but also solves the problem of vanishing gradient in the deep network, making the network easier to converge. The multi-branch structure increases the number of parameters to be computed and acquires more feature representations, but multiple branches mean that all branches are computed before the next step of fusion, which leads to the inability to make full use of the computational power of the hardware, increasing the amount of computation and reducing the speed. Therefore, although the multi-branch model brings high performance of the deep network, it becomes slower and occupies too much memory, so it is not suitable for applications in industrial scenarios. To solve these problems, RepVGG converts the multi-branch into a single path model in the inference stage and applies the idea of structural reparameterization to perform Op fusion and Op substitution, and Figure 4.2 shows the conversion process of the single path model.

The main process of step 1 in Fig. 4.2 is to fuse the convolutional layers in the residual blocks of each branch with the BN layer and the equivalent replacement convolutional layers of the identity layer. The red box in the figure indicates the fusion of Conv_3×3 with the BN layer, and the yellow box indicates the fusion of Conv_1×1 with the BN layer, the merging of layers can effectively improve the performance, and the fusion process is as follows:

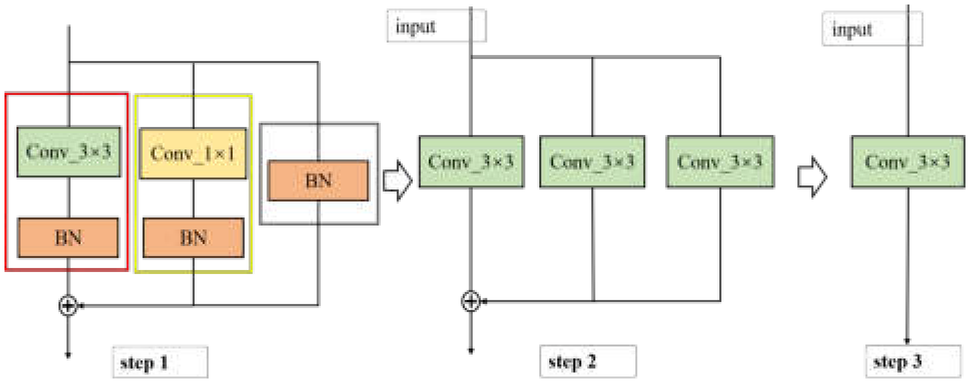


Fig. 4.2: Structure reparameterisation process.

Convolutional Layer Formulation:

$$Conv\ (\ x\ ) = w * x + b \tag{4.1}$$

BN layer formula:

$$BN\ (\ x\ ) = \gamma * \frac{x - \mu}{\sqrt{(\sigma)^2 + \epsilon}} + \beta \tag{4.2}$$

Among them, x is an element in the input feature map, w is the convolutional layer parameter, b is the bias term parameter, μ is the sliding mean of the BN layer, σ^2 is the sliding variance of the BN layer, γ and β are the scale factor and offset factor obtained from training and learning, and ϵ represents a minimal constant to avoid the denominator being zero. Since the BN layer is usually located after the convolutional layer, the output of the convolutional layer is used as the input parameter for the BN layer. Substituting (4.1) into (4.2) yields:

$$BN\ (\ Conv\ (\ x\ )\ ) = \gamma * \frac{w * x + b - \mu}{\sqrt{(\sigma)^2 + \epsilon}} + \beta \tag{4.3}$$

Further simplification leads to:

$$BN\ (\ Conv\ (\ x\ )\ ) = \frac{\gamma * w}{\sqrt{(\sigma)^2 + \epsilon}} * x + \left[\frac{\gamma * (b - \mu)}{\sqrt{(\sigma)^2 + \epsilon}} + \beta\right] \tag{4.4}$$

Re-order:

$$\begin{cases} \hat{w} = \frac{\gamma * w}{\sqrt{(\sigma)^2 + \epsilon}} \\ \hat{b} = \frac{\gamma * (b - \mu)}{\sqrt{(\sigma)^2 + \epsilon}} + \beta \end{cases} \tag{4.5}$$

Finally available:

$$BN\ (\ Conv\ (\ x\ )\ ) = \hat{w} * x + \hat{b} \tag{4.6}$$

After fusion, it is transformed into a convolutional layer, where $\hat{w}$ represents the weight parameters of the fused convolutional layer, and $\hat{b}$ represents the bias term parameters of the fused convolutional layer. The

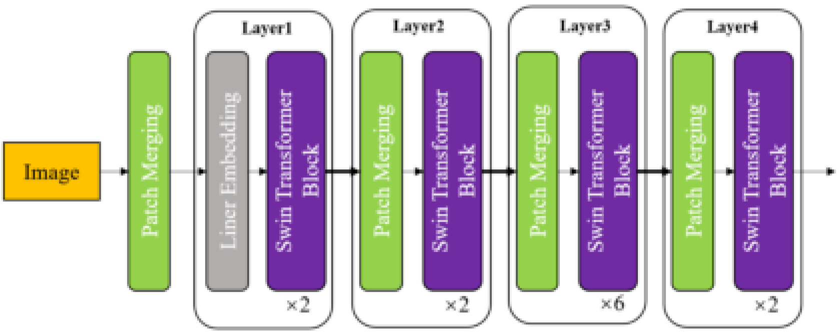


Fig. 4.3: Sliding window transform network structure.

entire process of fusing convolutional layers with BN layers does not increase computational complexity, but only modifies the convolution kernel to speed up the operation.

As shown in the yellow and grey boxes in Fig. 3, there are Conv_1×1 branches and identity branches in addition to the Conv_3×3 branch after fusion. Step 2 is to convert all the convolutions of different convolution kernels of other branches into Conv_3×3. The conversion of Conv_1×1 is mainly to use a matrix transformation to move the values in its convolution kernel to the center of Conv_3×3. Step 3 is to superimpose the convolutional weights and bias parameters of the three branches by using the same convolutional additive principle, and finally, the three branches are fused into a brand new Conv_3×3 single-path network structure.

RepVGG is more capable of enhancing the network's ability to express image features than ordinary convolutional layers for cases such as building construction environment occlusion and small targets at long distances. Worker helmet-wearing monitoring requires RepVGG to improve the detection accuracy while ensuring the detection speed to better adapt to the real-time monitoring needs. In this chapter, RepVGG is used to replace the 3×3 convolution in the Backbone, Neck, and Head parts to achieve the improvement of the overall performance of the network.

4.1. Introduction of the Swin Transformer Structure. Swin Transformer (Shifted Window Transformer) [15], a sliding window transform network, is a deep learning model designed for image recognition that improves on the Transformer and currently achieves state-of-the-art performance in computer target detection and image segmentation tasks. The Swin The general structure of the Transformer is shown in Figure 4.3.

To be closer to the original feature information, this paper chooses to introduce the Swin Transformer structure to replace the three CSPLayer layers similar to the Head in the PANet enhancement network. Compared with the CSPLayer structure, the Swin Transformer layer structure design can easily adjust the depth of the network, expand the sensory field of the network, extract features at different levels in the image, reduce the complexity of the entire PANet reinforcement network, and ultimately improve the target detection efficiency and accuracy, which is also beneficial for industrial real-time object detection or large-scale object detection tasks. This is also beneficial for industrial real-time object detection or large-scale object detection tasks. The improved PANet network with the introduction of Swin Transformer structure is shown in Figure 4.4.

4.2. Further decoupling of the detection header. YOLOX proposes a Decoupled Head for classification and regression tasks respectively, and the two branches are trained independently to achieve greater parallelization and faster convergence. Decoupled Head and Coupled Head have been compared and validated on the COCO dataset, and the accuracy has been significantly improved (see Figure 4.5).

The network can learn deeper features of the target object in the image, further improving the generalization ability of the network. In the helmet-wearing detection task, to further strengthen the network's ability to extract features and improve the network's generalization ability and robustness, this paper decouples the regression branch again and again to balance the positioning accuracy and classification accuracy. Figure 4.6 shows the comparison of the detection head before and after the improvement.

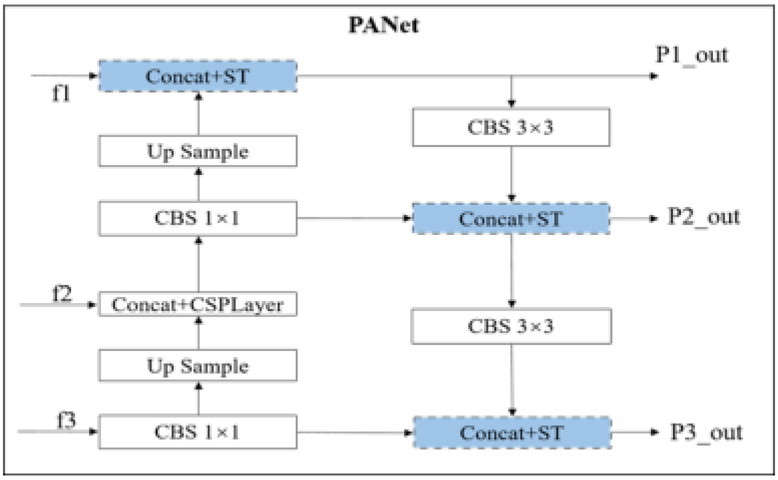


Fig. 4.4: Improved PANet network.

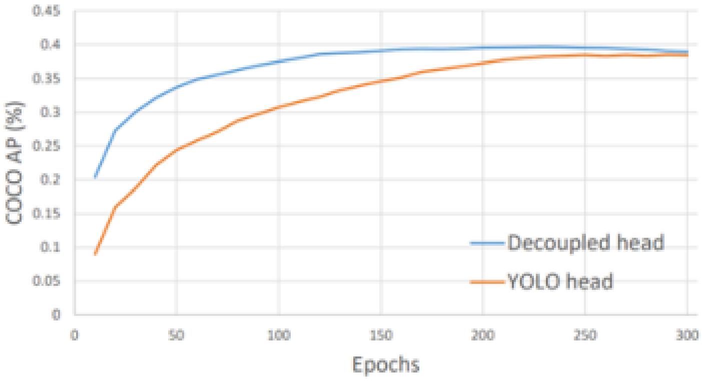


Fig. 4.5: Comparison of detection accuracyDecoupled Head.

4.3. Improved YOLOX network structure. In this paper, we improve on the YOLOX network structure by using the RepVGG structure parameterized model to replace the 3×3 convolutional layers of the three main parts of the Backbone, Neck, and Head, in which the activation function still uses SiLU (Sigmoid Linear Unit); in the PANet network, the Swin Transformer structure is introduced to replace the three CSPLayer layers; for the regression branch of the Decoupled Head detection head, the branch is further decoupled to independently perform the classification task, localization task, and confidence task. The overall network structure of the improved YOLOX is shown in Figure 4.7.

5. Construction of the dataset.

5.1. Data collection and labelling. To construct a dataset of workers wearing helmets in construction environments, images need to be collected, labeled, and numbered so that they meet the detection requirements of the model. One part of the images come from searching the keywords ”construction site helmet wearing” and ”construction workers” on the web, as well as images intercepted from construction videos of construction sites; the other part of the images are obtained by filtering and integrating open-source VOC and other public datasets on the web. publicly available datasets. Finally, these two parts of image data are merged and the detected objects in the images are labeled using the appropriate software. The annotation information of each image will generate a corresponding XML file so that it can meet the requirements of subsequent helmet model

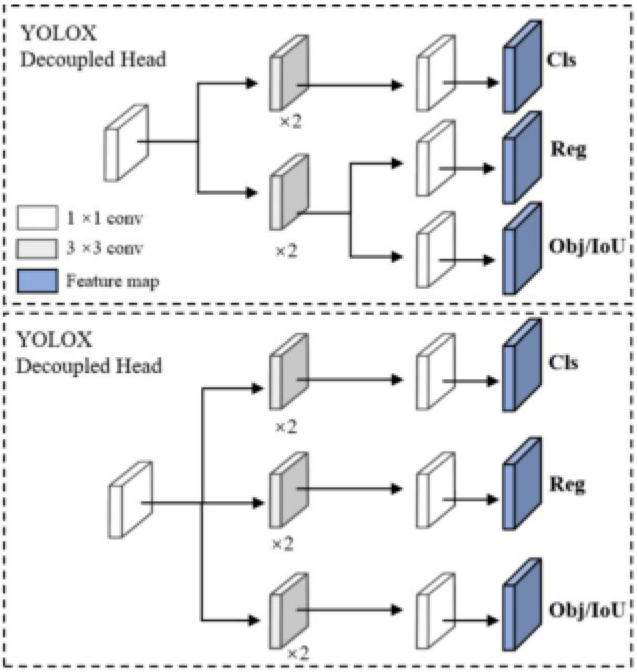


Fig. 4.6: Decoupled Head re-decoupled comparison.

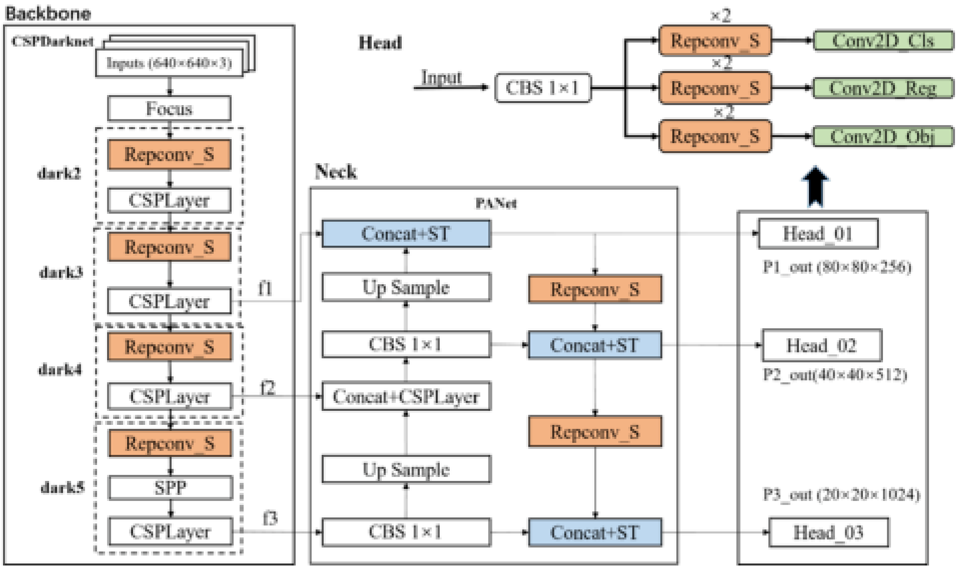


Fig. 4.7: Structure of the improved YOLOX.

training.

In this process, a total of 5973 images are collected, and the dataset is named ”safe_hat”. In the dataset, the original images are stored in the ”JPEGImages” folder, and their corresponding XML files are stored in the ”Annotations” folder. Some of the images in the dataset are shown in Figure 5.1.

In deep learning, the training of the model requires the division of the dataset images, which is generally

Fig. 5.1: Example of a partial image of the dataset.

divided into training set, validation set, and test set. In this paper, 5973 images in the dataset are randomly assigned to the training set and validation set according to the ratio of 8:2, and the test set is not set here because the validation set is not involved in the training, and at the same time can be used as a test set.

5.2. Data Enhancement. In YOLOX, both Mixup and Mosaic data enhancement methods are performed when the dataset is read. The Mixup method i.e., the obfuscation enhancement technique, uses the mixup function to linearly interpolate two different samples in the training dataset to mix them to generate new samples. Specifically for two samples, linear interpolation can be performed on their images, bounding box sizes, and categories respectively. Mosaic method i.e. mosaic enhancement technique, which stitches together four different images to form a new large image, and then randomly crops and scales this large image. Both methods can increase the diversity of the training data in the process of augmenting the data, thus improving the generalization ability and robustness of the model and reducing the risk of overfitting. In addition to these two methods, the model will use other methods for image data enhancement by each part of the structure during the training process.

5.3. Evaluation indicators. In object recognition for target detection, each detection frame can be regarded as a binary classification problem, where positive samples indicate that the detection frame correctly detected the target object and negative samples indicate that the detection frame did not correctly detect the target object. According to the classification results of the detection frame and the actual situation, it can be classified into the following four types:

True Positive (TP): i.e., the number of true cases, (generally set to 0.5) of the detection frame;

False Positive (FP): i.e., the number of detection frames of the false positive example (containing other redundant detection frames of the same true frame);

False Negative (FN): false negative, the number of undetected real boxes;

True Negative (TN): that is, the true negative case, the actual negative samples detected as negative samples.

(1) IOU i.e. Intersection Ratio. It is the ratio of the overlap area of the predicted and labeled boxes to their merged area. The larger the ratio, the more accurate the localization of the target. The image schematic is shown in Figure 5.2.

(2) Precision: i.e. the precision rate. It is the probability that the prediction is a positive sample and the actual sample is also positive, the formula is shown in equation (5.1).

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

(3) Recall. It is the probability of being predicted as a positive sample in a sample that is positive, generally the higher the recall, the lower the accuracy, the formula is shown in equation (5.2).

$$Recall = \frac{TN}{TP + FN} \tag{5.2}$$

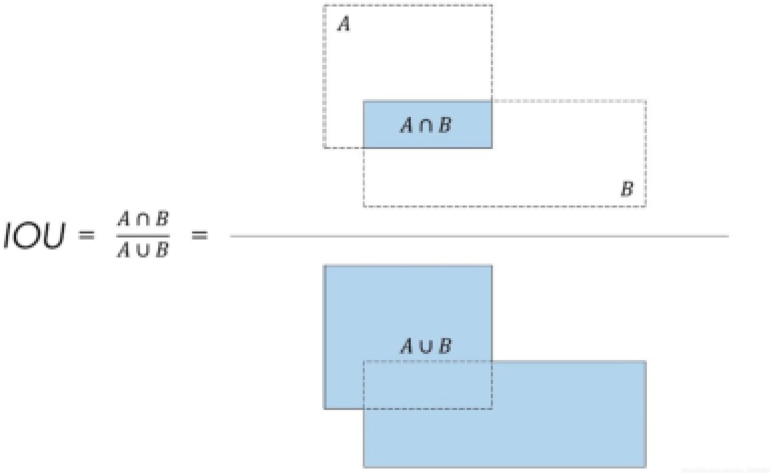


Fig. 5.2: Example of intersection-parallel ratio.

(4) AP: i.e. average precision. It is the area of the P-R curve with and as coordinates of the horizontal and vertical axes.

(5) mAP: i.e., mean average precision. It is the average value of all categories. It is a comprehensive measure to evaluate the comprehensive performance of the whole testing process by considering the two indicators of precision and recall.

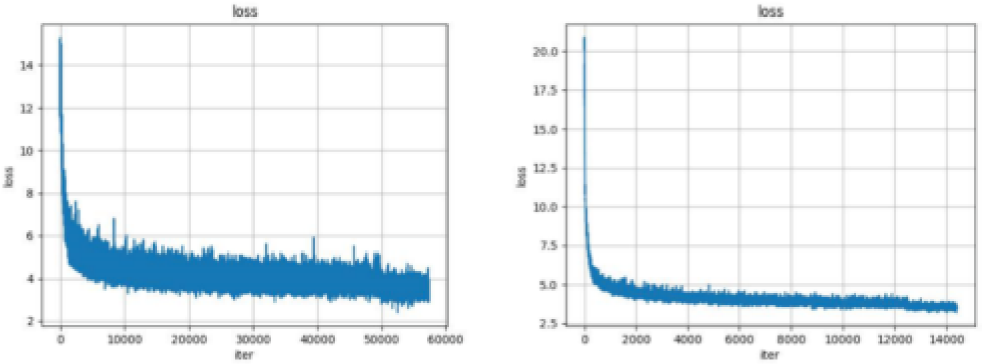
6. Calculation and Analysis.

6.1. Computer Operating Environment. This work is computationally intensive and the algorithmic procedures are mainly accelerated on GPU. The system environment is Ubuntu16.04, the GPU is NVIDIA Tesla V100, 32GB of RAM, PyTorch is used to build the deep learning framework, Python3.6 is used as the programming language, and some of the other acceleration tool libraries are Cuda10.5, Cudnn and so on.

6.2. Calculation Process and Analysis of Results. At the beginning of the training of this model, to make the model better adapt to the dataset, we choose to set the initial learning rate (learning rate) to 0.001, and the value of batch-size (batch-size) to 10, i.e., 10 images samples are selected for each training. The model uses the Adam optimizer, the learning rate decay coefficient is set to 0.9, the weight-decay coefficient (weight-decay) is set to 0.0005, and the confidence threshold is set to 0.5. During the experimental process, iteratively, the learning rate is gradually adjusted to 0.05, and the maximum number of training rounds (max-epoch) is 120, which is reached when the training is automatically stopped. The whole process is validated after the completion of each round of training, dynamically adjusting the model parameters and optimizing the model to avoid overfitting the model on the training set.

During the training process, the value of the loss function is used to present the model as good or bad, and the smaller the loss value, the better the model training. As shown in Figure 6.1, (a) graph represents the convergence process of the loss function during the training process of the original YOLOX model, and (b) graph represents the convergence process of the loss function during the training process of the improved YOLOX model. The horizontal and vertical coordinate values indicate the number of training iterations and the value of the loss function at that number of iterations, respectively.

From Figure 6.1, it can be seen that the improved YOLOX network model performs better with faster convergence and lower loss function values during training. To further evaluate the detection performance of the model, the next step is to analyze the detection effect of the model through the analysis of the P-R curve, in which the performance can be evaluated by the area (AP) enclosed by the curve and the coordinate axis, and the larger the area, the better the performance. Figure 6.2 exhibits the P-R curves obtained from the validation of the YOLOX model on the self-constructed helmet dataset before and after the improvement. As can be seen



(a) (b)

Fig. 6.1: Convergence plot of loss function before and after improvement.

Table 6.1: Comparison of helmet detection results of the model before and after improvement

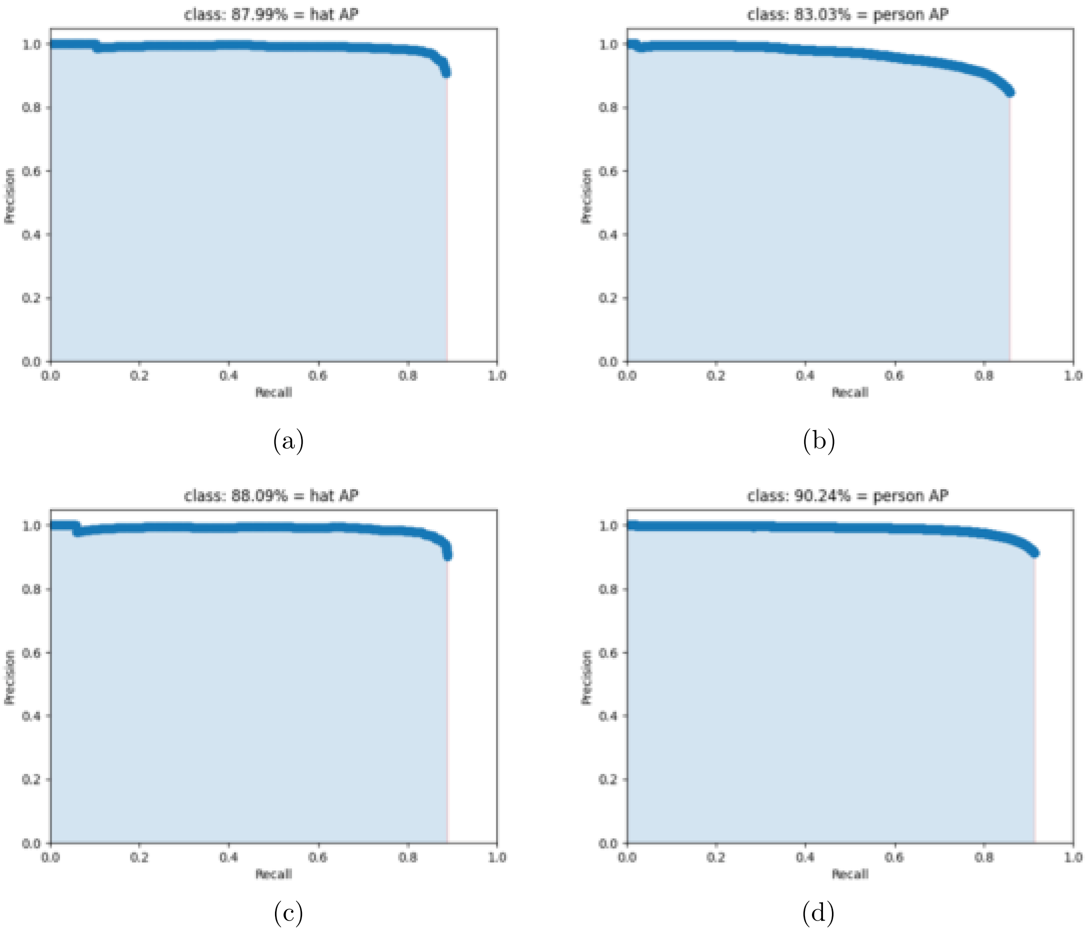
Algorithmic model	categories	Pre(%)	Recall(%)	mAP(%)	FPS
previous approach:	hat	94.67	87.36	85.56	10
	person	89.47	81.95		
our approach:	hat	94.78	87.77	89.12	11
	person	95.12	87.13		

from the figures, (a) and (b) show the P-R curves of the pre-improved YOLOX model for the categories ”hat” and ”person”, and (c) and (d) show the P-R curves of the improved YOLOX model for the categories ”hat” and ”person”, and (e) and (f) show the P-R curves of the improved YOLOX model for the categories ”hat” and ”person”. (c) and (d) show the P-R curves of the improved YOLOX model for the categories ”hat” and ”person”. The P-R curves of the improved model for both helmet wearers and non-helmet wearers are more biased towards the upper right corner of the coordinate axis and enclose a larger area. From the graphs, it seems that the improved model has improved the detection accuracy for both ”hat” and ”person” categories. Although the improvement in detection accuracy is small for the case of wearing a helmet, the improvement in detection accuracy for the case of not wearing a helmet is 7.21%, which indicates that the detection accuracy of the improved model has been significantly improved.

The P-R curve shows the goodness of the category detection results, and the final measure of the comprehensive performance of the model, i.e., the goodness of the multiple categories, is still based on the value of mAP. Table 6.1 shows the comparison of the results of the YOLOX algorithm model before and after the improvement after experimentation on the self-constructed helmet dataset.

The results showed that the improved YOLOX model achieved better detection results on the helmet dataset compared to the original YOLOX model. The improved model has a mAP value of 89.12%, which is a 3.56% improvement over the original YOLOX model. In terms of real-time performance, the number of images detected per second by the improved network also increased. This indicates that the improved model can more accurately and quickly detect whether a worker is wearing a helmet or not, and is more suitable for target monitoring tasks in real construction scenarios.

7. Conclusion. In this study, the YOLOX algorithm is used as the basic detection model to achieve automatic identification of construction workers not wearing helmets on supervised construction sites, and the performance of the algorithm is improved and optimized. The improved YOLOX network model improved



(a) (b) (c) (d)

Fig. 6.2: P-R curves for each category before and after improvement.

the detection accuracy of both non-wearing and helmet-wearing personnel over the pre-improved model, with a total mAP improvement of 3.56%. This study demonstrates that the enhanced YOLOX network model exhibits superior performance and reliability in detecting helmet-wearing at construction sites, showcasing significant practical value and potential for widespread application.

REFERENCES

[1] MOHAMMADFAM I,GHASEMI F,KALATPOUR O, et, *al.Constructing a bayesian network model for improving safety behavior of employees at workplaces[J]*,Applied Ergonomics,2017,58:35- 47.

[2] Golovanov R, Vorotnev D, Kalina D. , *Combining Hand Detection and Gesture Recognition Algorithms for Minimizing Computational Cost[A]*, 2020 22th International Conference on Digital Signal Processing and its Applications (DSPA)[C]. Moscow, Russia: IEEE, 2020: 1–4.

[3] LIU Y, JIANG W, , *ARTIFICIAL I.Detection of wearing safety helmet for workers based on YOLOv4[J]*, International Conference on Computer Engineering,2021:83-87.

[4] FANG Q, LI H, LUO X, et, *al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos[J]*, Automation in construction, 2018, 85: 1-9.

[5] LI Y G,WEI H,HAN Z,et, LU-*al.Deep learning-based safety helmet detection in engineering management based on convolutional neural networks[J]*, Advances in Civil Engineering,2020(6):1-10.

[6] WANG W,LI Y T,ZOU T,et, *al.A novel image classification approach via dense-mobilenet models[J]*, Mobile Information Systems,2020:1-8.

[7] Bochkovskiy A, Wang C Y, Liao H Y M., *YOLOv4: optimal speed and accuracy of object detection [EB/OL]*, (2020-4-23) [2023-5- 29].

[8] REDMON J, FARHADI A., *YOLOV3: An Incremental Improvement[EB/OL]*, [2022-03-23]. https://arxiv.org/pdf/1804.02767.pdf.

[9] He K, Gkioxari G, Dollár P, et, *al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision*, 2017: 2961-2969.

[10] Liu W, Anguelov D, Erhan D, et, *al. Ssd: Single shot multibox detector[C]*,//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[11] GONG M,WANG D,ZHAO X,et, *al.A review of nonmaximum suppression algorithms for deep learning target detection[C]*, //Seventh Symposium on Novel Photoelectronic Detection Technology and Application,2021.

[12] WANG C Y,LIAO H Y,WU Y H,et, *al.CSPNet:A new backbone that can enhance learning capability of CNN[C]*,Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,New York:IEEE,2020:390-391.

[13] Ge Z, Liu S, Wang F, et, *al. Yolox: Exceeding yolo series in 2021[J]*, arXiv preprint arXiv:2107.08430, 2021.

[14] Ding X, Zhang X, Ma N, et al., *Repvgg: Making vgg-style convnets great again[C]*, //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13733-13742.

[15] Liu Z, Lin Y, Cao Y, et, *al. Swin transformer: Hierarchical vision transformer using shifted windows[C]*, //Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jul 18, 2024
Accepted: Aug 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1348–1362, DOI 10.12694/scpe.v26i3.4314

EXPLORING A NEW MODEL OF COLLEGE ENGLISH TRANSLATION CLASSROOM VIA NATURAL LANGUAGE PROCESSING AND COMMUNICATION TECHNOLOGY

YUCHEN GUO*

Abstract. The crucial duty of developing translation skills for China's modernization falls on higher education institutions that teach translation. Information and intelligent technology are becoming increasingly ingrained in people's lives as civilization grows and develops. In this work, we use natural language processing and communication technologies to build a new type of university English translation classroom. To address the challenge of inferring semantic implication linkages in natural language processing, we put forth a deep learning model based on semantic rounding and semantic fusing. The technique can be applied to university translation classes to help basic translation tasks with effective reading comprehension. Furthermore, we developed a wireless classroom interaction system that enables effective interoperability between teachers and students in the classroom by embedding a natural language processing model in real time. Our natural language processing model performs exceptionally well and is capable of making predictions in real time, according to experimental results. The entire solution gives universities English translation classes a whole new experience.

Key words: Higher education,English translation, Interaction system, Wireless communication, Natural language processing model

1. Introduction. Classroom interactive systems have drawn increasing attention as a crucial tool to support classroom information interaction and assist teachers in understanding students' learning status in real time, given the ongoing development of education informatization and the ongoing transformation of traditional teaching methods [1]. As a significant area of study in computer science and artificial intelligence, natural language processing (NLP) has emerged in recent years. Its goal is to develop ideas and techniques that would enable humans and computers to communicate naturally. In addition to being a branch of computer science, natural language processing incorporates knowledge from other academic fields including linguistics and mathematics [2]. Although it focuses on human language in everyday conversation, its study is essentially distinct from that of traditional human linguistics. Instead of studying human language per se, natural language processing focuses on creating computer systems—particularly software systems—that enable efficient natural language communication between people and machines [3].

When it comes to teaching translation and classroom interaction systems, natural language processing is especially crucial. Teaching translation requires the capacity to comprehend and process material, and NLP offers several useful tools to help with this process. A "fill-in-the-blank" question, for instance, is comparable to a Cloze-style query in that the computer reads and comprehends the text, then extracts words or entities from the sentences and provides an answer in accordance with the query [4, 5]. Conventional models often encode the question and document, output the answer, or iteratively update the multilayer network's attention mechanism's focus of attention, finally producing the answer consisting of the words that have received the greatest attention. These methods, nevertheless, frequently overlook the larger context in favor of concentrating solely on a few chosen phrases [6]. This chapter focuses on the functions of semantic rounding networks and semantically aligned fusion methods in the construction of a new deep learning model, SDF-NN. We base this on the proposal of semantic rounding networks and semantic fusion methods. These two significant enhancements increase the accuracy of the model, encourage the complete fusion of local inference results, and lessen the detrimental effects of interfering semantics on the final prediction outcomes [7, 8].

Furthermore, we created a wireless classroom interaction system that supports real-time teacher-student interaction through the natural language processing model integration. In this way, natural language processing

*School of Foreign Languages Fuyang Normal University, Fuyang Anhui 236000, China (jiliangxie@jou.edu.cn)

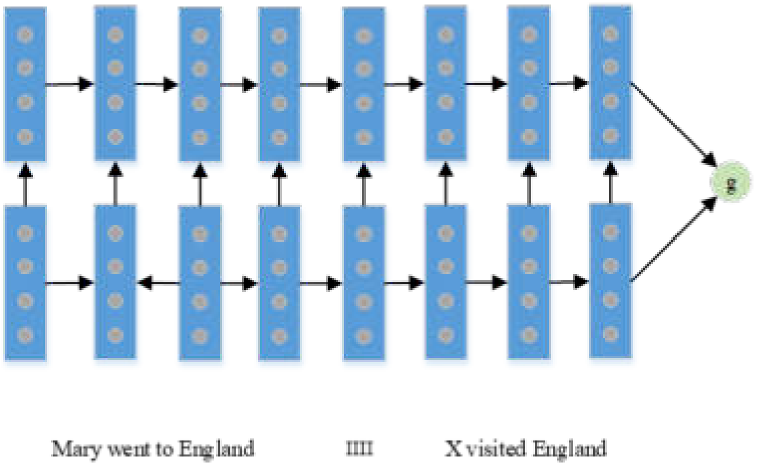


Fig. 2.1: Deep LSTM Reader Model.

strengthens the case for the modernized educational model by improving classroom interactions while also increasing the effectiveness of translation instruction.

2. Related work.

2.1. Natural Language Processing Based On Deep Learning.

1. Natural language understanding and in-depth education. An overview of the evolution of Natural Language Processing (NLP) with deep learning may be found in this section [9, 10]. It goes into great length on how machine learning challenges can be derived from problems with natural language understanding, including the requirement to convert computer-readable mathematical symbols from human-readable characters.

2. Inserting Words. The notion of word embedding and its significance in deep learning can be thoroughly explained in this section. First, the original word representation—one-hot representation—as well as its drawbacks—particularly the lexical gap phenomenon—are covered. After that, the idea of distributed representation (DR) is presented in order to clarify how it improves upon the One-hot representation's drawbacks [11, 12].

3. Vector bases of pre-trained words. Word vector libraries that have already been trained, like GloVe, can be covered in this area along with their uses and benefits for various NLP tasks (including named entity identification, word analogies, and word similarity) [13, 14].

Readers will be better able to comprehend the essence of each topic and its place in your research, and the sections on related work will be more logically organized.

2.2. Attentional Models. The first deep learning model that has been proposed, called Deep LSTM Reader, is a very basic model that encodes a query and a document independently using a two-layer LSTM, and then classifies data based on the representation that has been created from the two layers. This extremely basic model, which solely relies on encoding, misses the relevant characteristics of documents and queries. (Fig. 2.1).

The report then presented two models that Google DeepMind has developed: the eager reader model and the attentive reader model. First, the document and query are represented separately in the attention reader model. The query is encoded using bidirectional long short term memory (LSTM) and its overall representation is obtained by splicing the outputs of the forward and reverse last hidden layers. Similarly, the document is also encoded using bidirectional LSTM and its encoded representation of each lexical element (i.e., word in the document) is obtained by splicing the corresponding forward and reverse hidden layers. Finally, the overall representation of the document is a weighted average of all the lexical elements in the document, where the

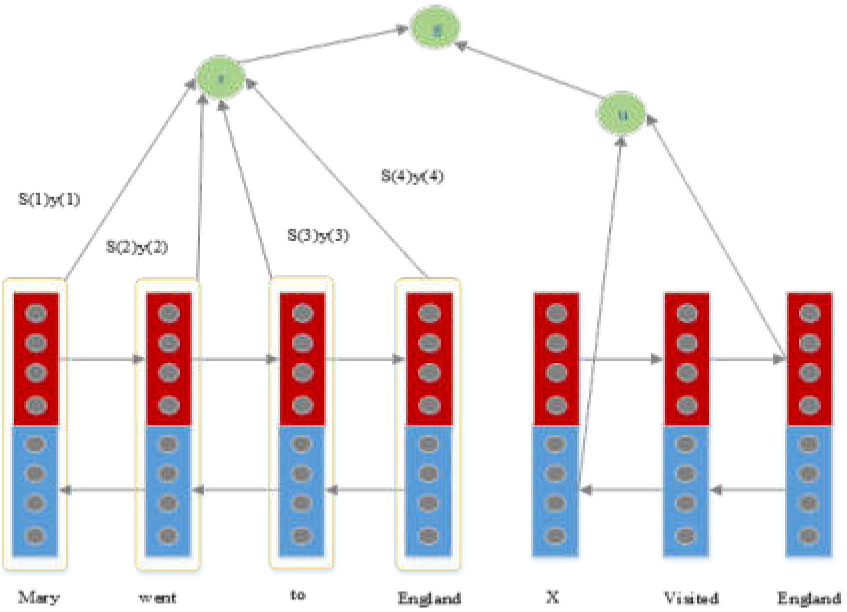


Fig. 2.2: Attentive reader model.

weights are determined by the generated attention, and the weights indicate the importance of the corresponding lexical elements in answering the query.

These models have a direct connection to the classroom interaction system and model you have suggested. Similar attentional mechanisms might be employed in the system you are building to gauge how important various text segments are in responding to students' inquiries and maximizing teacher-student engagement. Additionally, when handling bidirectional information flow in classroom interactions, the deployment of bidirectional LSTMs may also be instructive. This implies that important technological components from these linked works may immediately offer your suggested system technical help and inspiration.

Then the documents and query representations are used for classification [15, 16]. This is shown in Fig. 2.2. The Impatient Reader has been improved in that instead of encoding queries as a whole as in the Attentive Reader model, the token of each query is related to the token of the document. This mechanism is similar to reading each token in a query and then focusing on the information of the corresponding token in the document [17]. This model has a more complex attention mechanism, but it may not be effective, because in terms of the actual human reading comprehension mindset, it is impossible to read a word in the query and then go back to read the original text again when answering the question, which is too inefficient, and long documents may also affect memory. The structure of the model is shown in Fig. 2.3.

The Stanford AR model mainly uses a bidirectional LSTM to encode the document and query separately, and uses the correlation between the words and the query to obtain the Attention value, which is used to weight the embedding of the document to obtain a final output vector for answer prediction. The model structure is shown in Fig. 2.4.

The Attention Sum Reader model obtains the associated representation vectors by encoding the document and the question, respectively, using two bidirectional GRUs. The outcome, which can be seen as the attention matrix, can be thought of as the weight of each token in the document in relation to the query. Ultimately, each token's likelihood in the text is normalized using the softmax function, and the result with the highest probability is regarded as the answer to the question. That's what the figure depicts. The AOA model is suggested because Q&A should be predicated on the mutual attention of the inquiry and the document, whereas the aforementioned models are based on one-way attention.

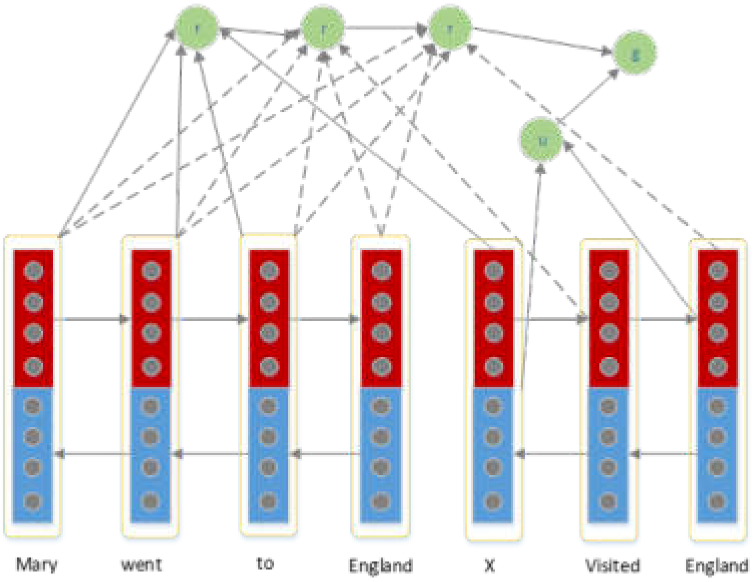


Fig. 2.3: The model of the patient reader.

2.3. Classroom Interaction. The so-called intelligent classroom is actually a multimedia classroom that operates and controls audio-visual equipment, computers, projectors, light, electricity and other devices in the classroom, facilitates access to teaching resources and teaching activities for teachers and students, and provides information storage and real-time feedback [18]. Another development idea of classroom interactive system is to modify and upgrade the existing classroom hardware in a limited way. Both of these ideas can facilitate interaction between teachers and students, but they also have many differences [19]. There are many educational research institutions and companies in China and abroad that are focusing on smart classrooms, including McGill University, the University of Chicago, DELL, and Intel Corporation. For example, DELL has proposed an intelligent classroom solution with the goal of creating an interactive and collaborative learning environment.

In China, some companies have also launched intelligent classroom solutions, such as the Xunjie II intelligent classroom developed by Shanghai Excellence Electronics Co. The intelligent classroom mainly consists of the following components: control panel, fully automatic guide, multimedia external devices, teacher and student scene cameras, etc. Users can choose student interaction modules according to actual needs, as shown in Fig. 2.5.

At present, research on smart classrooms has made some progress in both theoretical and applied research, and these devices have met the needs of teachers and students for daily interaction to a certain extent [20].But these gadgets necessitate a significant hardware upgrade for the classroom, which is expensive in terms of engineering, time, and renovation expenses. Upgrading any component of the system later on is likewise challenging. An alternative perspective on the teacher-student interaction system is comparable to classroom voting devices like the SunVote Conference Voting Device S52Plus, which can communicate at a distance of approximately 30 meters, uses 2.4G frequency wireless radio frequency technology, is moderately sized, and uses CR2032 coin cell batteries. These devices are used by many college students in North America. It is mostly utilized for staff training, yearly meetings, product roadshows, academic conferences, quality assessment and evaluation, etc.

3. Methodology.

3.1. Semantic Entailment Relation Inference Model Based On Semantic Discarding And Fusion. The relationship between the two sentences is incorrectly anticipated to be contradictory if the strongest

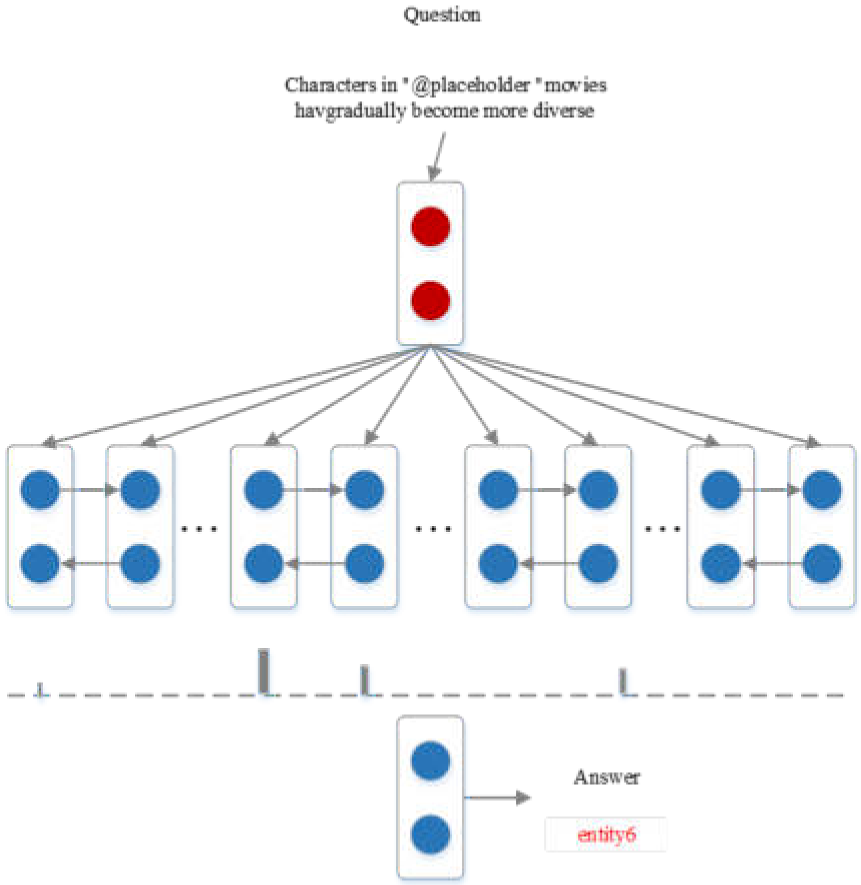

Question
Characters in "@placeholder "movies havgradually become more diverse
...
Answer
entity6

Fig. 2.4: Stanford ar model.

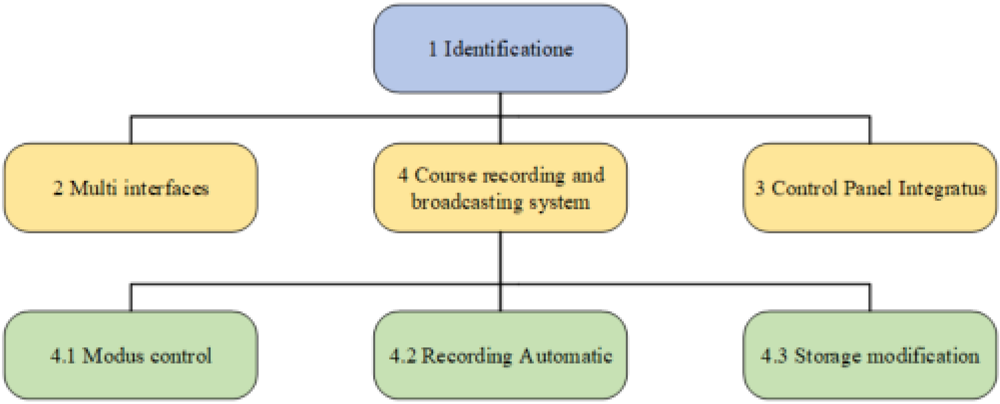


Fig. 2.5: Functional block diagram of excellent intelligent classroom.

semantic relation is utilized as the final prediction. To predict the proper result as "neutral," the model needs to take into account the combined findings of many local inferences. Therefore, in order to rationally fuse all

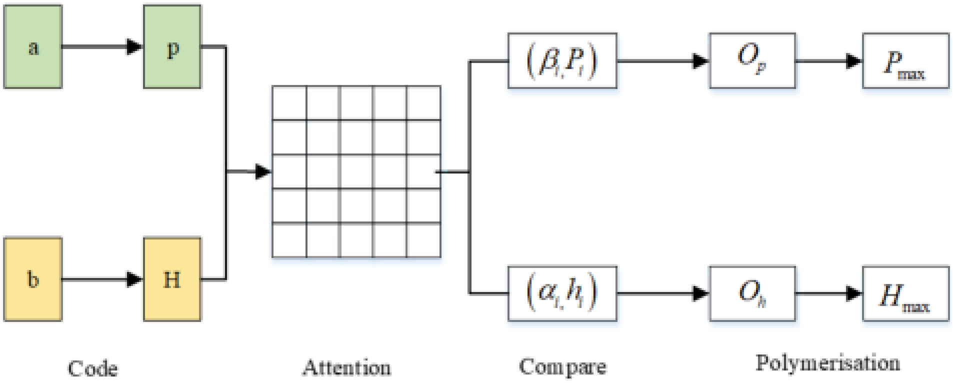


Fig. 3.1: Network structure of model based on decomposed attention.

local inference outcomes, this chapter suggests a semantic fusion alignment approach. Drawing from the aforementioned constraints of the two earlier models, the first step involves designing a semantic discarding network (SDN) to exclude extraneous or disruptive semantic information during the "comparison" phase. While all extracted semantic information is passed forward in this network, some semantic information is purposefully dropped during training.

In previous deep learning models based on decomposing attention, four processing steps are generally summarized: coding, attention mechanism, comparison, and aggregation. This is shown in Fig. 3.1.

Assume two sentences a and b, a is an Embedding representation of the premise of length m and b is a word vector representation of the hypothesis of length n. After encoding the two sentences using the encoder, the encoded representation of the word vector is obtained as the matrix $P = [P_1, \ldots, Pm], \forall i \in [1, \ldots m]$ and $H = [h_1, \ldots, hn], \forall i \in [1, \ldots n]$. The corresponding aligned text pairs are then obtained by applying the corresponding decomposition attention mechanism to the attention matrix. (α_j, h_j) and $(\beta_i, p_i)$are called aligned text pairs, where α_j is a subfragment in P that is aligned withh_j , which is a recoded representation of P based on the attention distribution of h_j to each word in P . β_i is a subfragment in H that is aligned with p_i, which is a recoded representation of H based on the attention distribution of pi to each word in H . This is also known as aligning related segments based on attention and highlighting salient features. The above steps are designed to extract semantic features from the sentences. In the "comparison" phase, a single feedforward neural network or LS TM network is usually used to extract the semantic features between aligned text pairs and obtain the corresponding local inference results . The aligned text pairs are fed into the feedforward neural network G, and the local inference result O is obtained as shown in Fig. 3.2. However, in natural language, different aligned text pairs have different relationships, and the internal feature relationships are not consistent, which means different local inference results. For example, "wears, dressed in" means the same thing, and "in the morning, at night" means the opposite. Therefore, different functions should be applied to extract the different relationships between the aligned text pairs, as shown in Fig. 3.3. Each aligned text pair is passed through k different feedforward neural networks G. Through the learning of these feedforward neural networks, multiple local features are generated for each aligned text pair. The gate function g is then defined to determine the weight of each G function, i.e., each local inference result, and finally weighted and processed so that each pair of aligned text pairs still produces a corresponding local inference result. This inference result will be more accurate.

As analyzed in the previous section, these traditional networks extract and analyze all the extracted semantic information, which includes interfering semantics that can have a negative impact on the final model. Therefore, to address this problem, we propose a feedforward discard network (SDN) that differs from the traditional approach, as shown in Fig. 3.4.

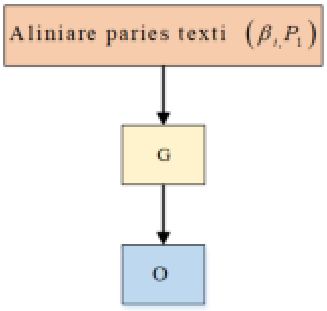


Fig. 3.2: Local inference results obtained through a feedforward neural network.

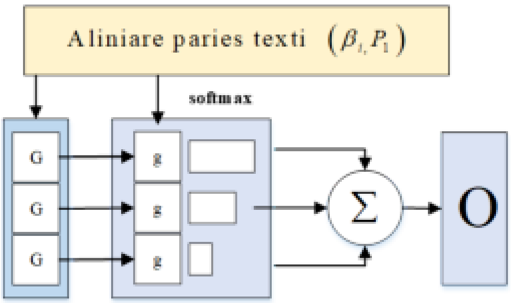


Fig. 3.3: Local inference results obtained through multiple feedforward neural networks.

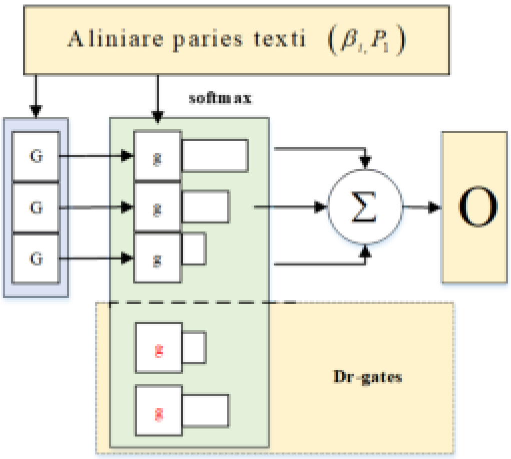


Fig. 3.4: SDN Structure.

3.2. Semantic Fusion Alignment Methods. In the traditional approach, the obtained local inference results are aggregated directly by maximum pooling or average pooling. This is shown in Fig. 3.5.

The obtained local inference results are directly processed by maximum pooling to the most significant features, and then fed into the softmax layer for the final prediction. This simple treatment, as analyzed above, ignores the relationship between the local inference results and does not include their combined decision information. Each column of the local inference outcome matrices Op and Oh is a local inference outcome, and

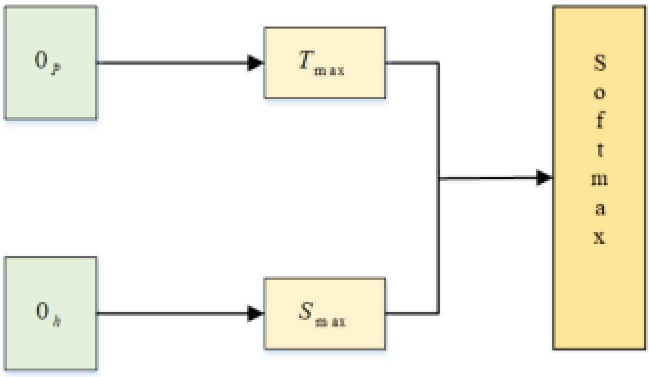


Fig. 3.5: Processing local inference results with max pooling.

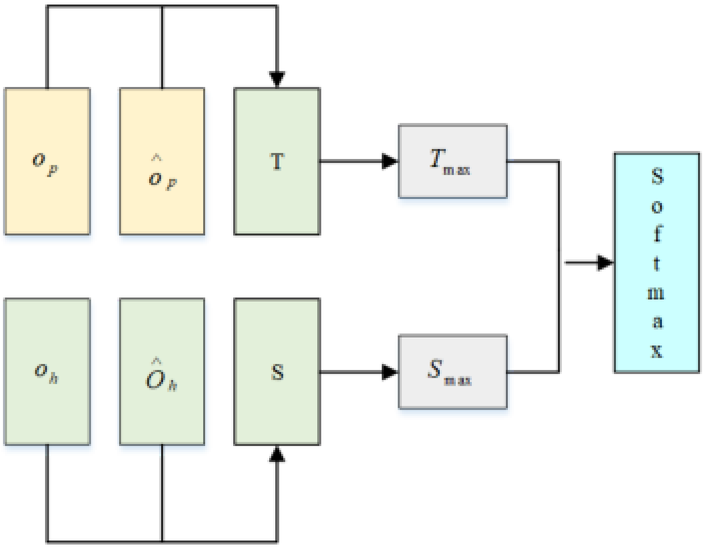


Fig. 3.6: Structure of SFA.

each row of their corresponding transpose matrices represents a local inference outcome.First we multiply the local inference result matrix with its transpose matrix (see Fig. 3.6):

$$U = O_p^T O_p \tag{3.1}$$

$$V = O_h^T O_h \tag{3.2}$$

U11 in the U matrix represents the attention of the first local inference result in O_p to itself, U21 represents the attention of the first local inference result in O_p to the second inference result, and the same for the remaining elements. each element of the V matrix is also aligned with the elements of the U matrix. The main diagonal elements of the resulting alignment matrices U and V are set to 0, because the elements should not be aligned with themselves and the attention of the local result on itself needs to be eliminated.

$$u_i = (U_i) \tag{3.3}$$

$$v_j = (V_j) \tag{3.4}$$

$$\hat{O}_p^i = O_p \cdot u_i, \forall i \in [1, \ldots, m] \tag{3.5}$$

$$\hat{O}_h^i = O_h \cdot v_j, \forall j \in [1, \ldots, n] \tag{3.6}$$

The elements in the U matrix are normalized by row. After normalization, the first column u1 of the matrix represents the attention weight distribution of the first partial inference result on the other inference results in O_p. The second column u2 represents the attention weight distribution of the second partial inference result on the other inference results in O_p . The second column u2 represents the attention weight distribution of the second partial inference result on the other inference results in O_p , and so on. After normalization, the first column of the matrix, v1, represents the attention weight distribution of the first inference result to the other inference results, and the rest is the same.

3.3. Overview of GIS technology and its related applications. We present a comprehensive inference model for semantic entailment relations based on the decomposed attention mechanism, SDF-NN (Semantic Dropping and Fusion Neural Network), based on the previous findings and the two novel networks and methodologies. We present a comprehensive inference model for semantic implication relations called SDF-NN (Semantic Dropping and Fusion Neural Network), which is based on the decompositional attention mechanism. This model can fully illustrate the efficacy of the semantic fusion alignment (SFA) and semantic dropping network (SDN) techniques. The model has the same four steps as the previous decomposition-based attention model: coding, attention, comparison, and aggregation. The overall framework of the model is shown in Fig. 3.7. Assume that two sentences$a = (a_{1,....,}a_m)$and $b = (b_{1,....,}b_n)$ a is the word vector representation of the premise of length m (Word Embedding representation) and b is the word vector representation of the hypothesis of length n. We assume that a_i , $b_i \in Rd$ and d is the dimension of the word vector. The goal is to predict the label y to determine the relationship between a and b. y can be Neutral, Contradiction, or Entailment. (1) Sentence encoding First, two sentences are input into the bidirectional LS T M in temporal order, and the output of each hidden layer is used as the encoded representation of the word vector of the corresponding input.

$$p_i = biLSTM\,(a_i) \tag{3.7}$$

$$h_j = biLSTM\,(b_j) \tag{3.8}$$

Matrix $p = [p1, \ldots, pm] \in R2r \times m$ and $H = [h1, \ldots, hn] \in R2r \times n$ is the output of boils T M hidden layer. r is the number of neurons in the hidden layer. the LSTM model is formulated as follows.

$$i_t = \sigma\,(W_i x_t + U_i h_{t-1}) \tag{3.9}$$

$$f_t = \sigma\,(W_f x_t + U_f h_{t-1}) \tag{3.10}$$

$$o_t = \sigma\,(W_o x_t + U_o h_{t-1}) \tag{3.11}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\,(W_c x_t + U_c h_{t-1}) \tag{3.12}$$

$$h_t = o_t \tanh\,(c_t) \tag{3.13}$$

$a_i$and b_j are entered into the model sequentially according to the time series t, respectively, as x_t , LSTM utilizes a memory version, including an input gate i_t , a forget gate ft , an output gateo_t , and a memory unit c_t , to generate a hidden layer output h_t . Also, using $biLSTM$ encoding is more efficient than $LSTM$ encoding. It is to encode the sentence forward and then encode it backward, and each word will have two corresponding hidden layer outputs, which will be concatenated as the final encoded form of the word.

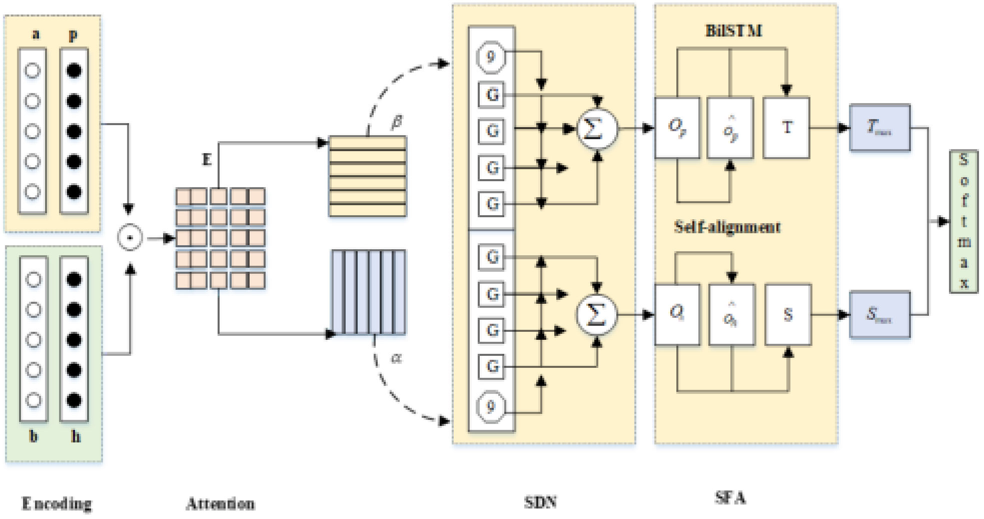


Fig. 3.7: SDF-NN Structure Diagram.

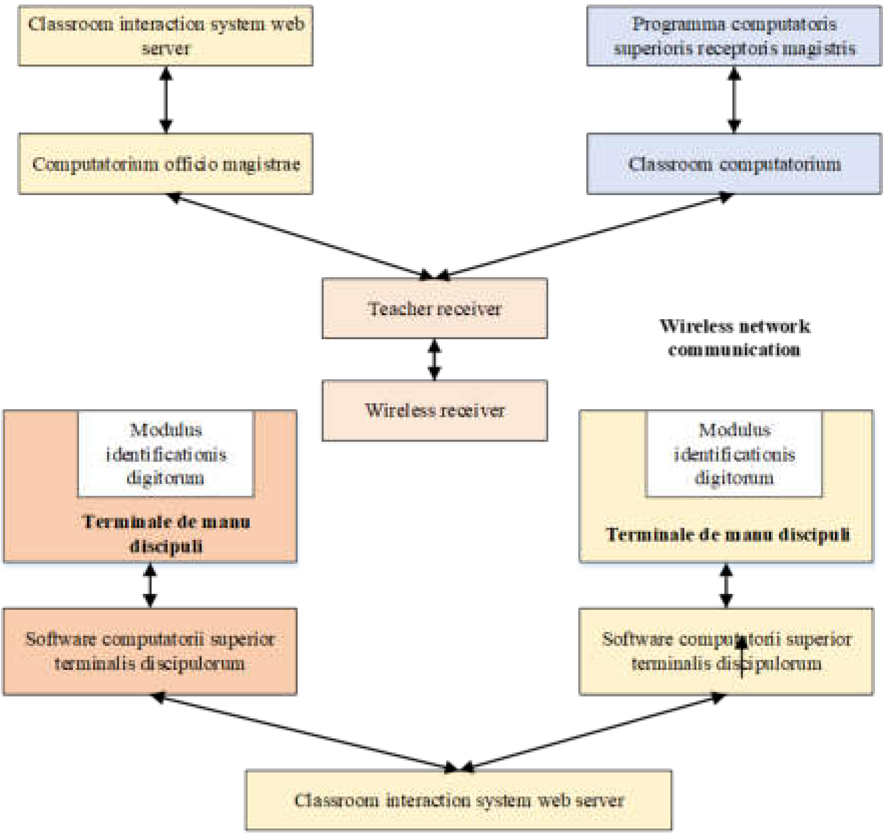

Classroom interaction system web server
Programma computatoris superioris receptoris magistris
Computatorium officio magistrae
Classroom computatorium
Teacher receiver
Wireless network communication
Wireless receiver
Modulus identificationis digitorum
Terminale de manu discipuli
Modulus identificationis digitorum
Terminale de manu discipuli
Software computatorii superior terminalis discipulorum
Software computatorii superior terminalis discipulorum
Classroom interaction system web server

Fig. 3.8: Schematic diagram of classroom interaction system based on wireless communication.

3.4. Functional Modules Of The Wireless Communication-Based Classroom Interactive System. Based on the needs of most users, this wireless communication-based classroom interaction system is designed. The specific structure of the wireless communication-based classroom interactive system is shown in Fig. 3.8.

The teacher receiver host software is installed on the classroom computer, and the teacher receiver and wireless receiver are fixed and physically connected in the classroom. To attend class, each teacher merely needs to place the unique SD card into the teacher terminal. The instructor just needs to bring their SD card with them after class to use the office card reader to transmit the data from the SD card to the web server. The main functions of the classroom interactive system are: (1) The webmaster first adds class and student and teacher user information. (2) The student user logs into the web page and connects the student handheld terminal to the web page through the student handheld terminal host software. Then the MAC address serial number of the handheld terminal chip is passed into the front-end of the webpage, followed by the fingerprint registration of the student handheld terminal, and the address serial number is bound to the user name after the registration is successful. If the user has already registered the fingerprint or the student's handheld terminal has been registered by other students, the above operation cannot be performed.(3) The teacher downloads the list of students and their corresponding MAC address numbers of the class he/she wants to teach from the web page and puts them into the SD card for the teacher's receiver. The SD card is then inserted into the Teacher Receiver during class. The software on the teacher's receiver has various function buttons to enable the class roll call, class quiz ABCD questions or Y/N questions, collect the students' answers and display them on the projector, and then store them on the SD card on the teacher's receiver. (4) Finally, the information stored in the SD card can be stored in the web server for subsequent viewing and processing, so that we can have a more comprehensive understanding of the actual learning status of each student.

4. Experiments.

4.1. Dataset And Parameter Settings. The model was assessed using the Stanford Natural Language Inference (SNLI) dataset, which is made available to the public. The 570k English sentence pairings in the dataset have been manually labeled by several people. Three sets make up the dataset: a test set, a validation set, and a training set. The model is trained using the training set, verified using the validation set, monitored during training to avoid overfitting, then tested once more using the test set to assess the model's performance. As in previous work, we remove the samples labeled with "-" (the "rest" class) from the dataset, leaving 549,367 sentence pairs in the training set, 9842 sentence pairs in the validation set, and 3,632 sentence pairs in the test set. The remaining samples in the training set are 549,367 sentence pairs, the remaining samples in the validation set are 9842 sentence pairs, and the remaining samples in the test set are 9824 sentence pairs. In this model, the dropout of the feedforward neural network in the SDN is set to 0.2, and the dropout of the rest of the model is set to 0.3. Set the training batch size to 128, i.e., the number of samples for one input model training is 128, and the loss function for this training is the average of these 128 samples. Depending on the performance of the training machine, this value can be adjusted accordingly. In the SDN layer, the final experimental model is set up with 3 G-functions and 5 gate functions, which can be tuned in more detail to achieve better performance. The training loss function is a multi-class cross-entropy function, and the optimization method is Adam (Adaptive Moment Estimation), which has the advantage that after bias correction, the learning rate of each iteration has a fixed range, making the parameter correction is relatively smooth. The whole model is implemented based on TensorFlow, a second-generation artificial intelligence learning system developed by Google based on DistBelief.

4.2. Performance Analysis. Table 4.1 shows the accuracy comparison of the SDF-NN model and some related models trained and tested on the SNLI dataset. Para is the number of parameters, the first row of the table is a classifier-based feature extraction model, which is considered as a benchmark comparison for the semantic implication inference problem. The next set of models (2) to (3) are based on sentence encoding. The third group of models (4) to (7) are based on attentional mechanisms. The last group of (9) and (11) models are integration models. It can be seen that the proposed model, SDF-NN, achieves the highest accuracy of 88.2% in a single model. At the end of the table, an elimination analysis is also performed to show the impact of two key design modules of the model on the overall model performance. We first remove the drop gates (dr-gates)

Table 4.1: Performance Comparison of Models on SNL Datasets.

Models	Para	Train(%)	Test(%)
(I)Unigram and bigram features [Bowman et al. 2015]	-	99.7	78.2
(2)300D LSTM encoders [Bowman et al, 2016]	3.0M	83.9	80.6
(3)300D Tree-based CNN encoders [Mou et al, 2015]	3.5M	83.3	82.1
(4)100D word-by-word attention [Rocktaschel et al, 2015]	250K	85.3	83.5
(5)600D BILSTM with intra-attemton [Liu et al. 2016]	2.8K	85.9	85.0
(6)200D decomposable attention models[Pankh et al, 2016]	580K	90.5	86.8
(7)300D re-read LSTM [Sha et al- 2016]	2.0K	90.7	87.5
(8)60OD ESIM[Chen et al. 2016]	4.3M	92.6	88.0
(9)600D ESIM+ Syntactic(Ensemble)	7.7M	93.5	88.6
(10)BIMPM [Wang et al, 2017]	-	-	88.9
(11)BIMPM(Ensemble)	6.4M	93.2	88.8
5OOD SDF	6.4M	92.8	88.2
5OOD SDF w/o dr-gates (SDN)	6.3M	91.0	87.7
500D SDF w/o SDNS	5.3M	90.3	87.5
5OOD SDF w/o SDN+SFA	1.5M	90.1	87.0

Table 4.2: Decomposition accuracy of SDF-NN model.

Models	N(%)	E(%)	C(%)
(Bowman et al. 2016)	80.6	88.2	85.5
(Wang ang Jiang 2015)	81.6	91.6	87.4
(Parikh et al. 2016)	83.7	92.1	86.7
SDF-NN(ours)	84.3	92.0	88.1

from the SDN network and the accuracy drops to 87.7%. Then we remove the entire SDN network and replace it with a simple feedforward neural network, and the accuracy drops to 87.5%. Finally, the SFA is removed and a simple maximum pooling method is used to handle those local inference results, and the accuracy drops to 87.0%. As can be seen, several key parts of the model are designed to have a very positive effect on the performance of the model.

Table 4.2 presents the accuracy results of the model for each of the three categories tested in the validation set of the SNLI dataset. It can be seen that the overall accuracy of the model is mainly due to the ”implicit” category, while the main accuracy loss is in the ”neutral” category. The reason for this may be that for the ”implicit” category, it is beneficial to discard the distracting information and consider the relationship between the segments globally for the final inference. However, for the ”neutral” category, there may not necessarily be a direct correlation between the segments of the statement, and forcing some segments to be aligned when decomposing attention may have a negative impact on the final result.

There are two very important parameters in the SDF-NN model, the number of functions G x and the number of discarded gates (drgates) y. Since the weight of all gate functions sums to 1, it is considered that the larger the number of y, the higher the weight of discards, i.e., the more information is discarded in the model. Fig. 4.1 shows the accuracy using different settings of the x and y parameters. Two patterns emerge: (a) First, we fix y to 2 and increase x from 0, and the accuracy rate starts to increase and then level off. This is because the function G serves to fully extract features from multiple aligned text pairs, i.e., local inference results, and we need enough functions G to extract feature information from the data, but this extraction ability will level off as the function G continues to increase. In this case, x is best set to 3. (b) Then, we fix x at 3 and increase y from 0 to 4. The model achieves the best performance for y = 2. This reflects that discarding information has a positive effect on the performance of the model, but too much information is discarded as y increases, which reduces the performance of the model. From these two sets of experiments, and based on the consideration

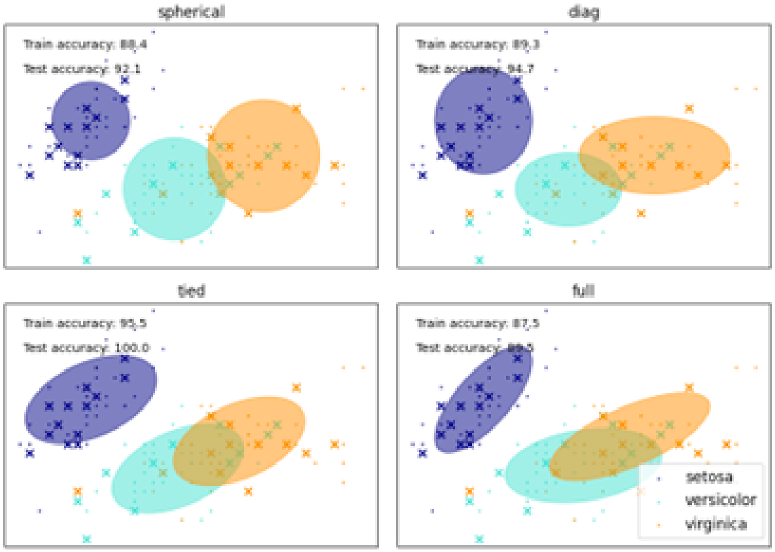


Fig. 4.1: Analysis of key parameters of SDF-NN.

Table 4.3: Comparison of Accuracy on Multinl I Dataset.

Models	SNLI (%)	Matched (%)	Mismatched (%)
Most frequent	34.3	36.5	35.6
CBOW	80.6	64.8	64.5
BILSTM	81.5	66.9	66.9
ESIM[Chen et al , 2016]	86.7	72.4	71.9
SDF-NN (ours)	88.2	77.5	76.5

of reducing the complexity of the model parameters, the optimal model parameters were determined to be 3 functions G and 2 discard gates (dr-gates) .

4.3. Performance Evaluation On The Multinli Corpus. In addition to the more widely used SNLI corpus for training, Stanford has recently released a new corpus, MultiNLI (Multi-Genre Natural Language Inference). The MultiNLI corpus contains 433k pairs of sentences that are annotated with textual implication information. This corpus is modeled after the structure of the SNLI corpus, but differs in that it covers more types of spoken and written language, and thus the dataset has greater diversity and complexity. In particular, the corpus contains samples from more than ten sources, including ten types of corpus from texts, spoken dramas, and so on. The validation and test sets contain samples from all ten types of corpora, but the training set contains only five types. The test data set of the corpus contains two additional types of data. Matched examples and Mismatched examples. Matched examples means that the sample type is present in the test set and also in the training set. Mismatched examples means that the sample type appears in the test set but not in the training set. The performance of the SDF-NN model can be evaluated on both test sample sets, and the relevant parameter settings for the experiments on this corpus are consistent with those on the SNLI corpus.

Table 4.3 shows the evaluation of the different models on this corpus. The first one is a CBOW model based on bag-of-words model and the second one is modeled with a basic bidirectional LS T M. ESIM is the base model without the tree LS T M structure. the SDF-NN model achieves an accuracy of 77.5% on the Matched test set and 76.5% on the Mismatched test set, both outperforming the other models. This indicates that the SDF-NN model has strong learning ability and generalization ability, and is suitable for more complex data sets.

This chapter proposes a memory mechanism for reading comprehension problems. Through the memory correction mechanism, the original document information is fused to correct the attention during the iterative learning process of the network, so that the original information is taken into account in each iteration of the attention, preventing significant attention bias and eventually outputting more accurate answers. Based on this memory mechanism, we designed the memory Gated Attention Reader (mGA) model, an end-to-end neural network-based model, to address the problem of attentional bias that occurs in models based on inference mechanisms. On the CNN dataset, Daily Mail dataset, and CBT data, we show that the prediction accuracy of our model is higher than that of some of the previously proposed models, and validate the effectiveness of the proposed memory correction mechanism. This memory mechanism is also different from previous models such as MemNet, which does not store the combined information of documents and queries in a separate component of the network, but directly introduces the original information into the network iterations repeatedly, solving the problem of information compression and loss caused by the deepening of the network. This approach is also a good analogy to the way the human brain works, where each time the focus of a query is found in an article, it is always based on a global article memory context. Future work will apply this idea of introducing raw information to solve the network information compression problem to other problems and propose a more general memory mechanism.

5. Conclusion. We propose a strategy to combine natural language processing and wireless communication for English classroom translation for university students. In this paper, we propose a semantic discard network and a semantic fusion alignment method for the semantic implication relation problem through the analysis of human thinking process of natural utterance relation determination, and propose an SDF-NN model, an end-to-end neural network model based on these two innovative methods. The SDF-NN model, an end-to-end neural network model, is proposed based on the two innovative methods. The natural utterances often have interference semantics on the final relation judgment, and the semantic discard network can discard the interference information to obtain more accurate local inference results, and then use the semantic fusion alignment method to align the relationship between the local inference results and better fuse these local inference results. The SDF-NN model achieves an accuracy of 88.2% in the public dataset SNLI, which is higher than the single model proposed by other related studies. The SDF-NN model also achieves 77.5% and 76.5% accuracy on the latest dataset, MultiNLI. This demonstrates that the model can learn more complex datasets and has the ability to learn generalization. In general, we provide a new model for the university English translation classroom.Future research can investigate the model's potential for use in other educational domains, such as multidisciplinary language learning and the creation of technological tools for translation. In addition, going over the model's flexibility and potential for generalization in other linguistic and cultural contexts will give readers a more thorough understanding of the model's applicability. Furthermore, it may be worthwhile to suggest avenues for further model modification, such as refining the algorithm to increase translation accuracy and real-time performance and assessing the model's efficacy in practical teaching situations.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

Funding Statement. The Planned Project of Fuyang Social Science Federation: Implement moral cultivation and improve the education in Fuyang City: Long-term mechanism of teacher ethics construction from the perspective of "the Balanced Development of Five-Education" (FSK2023026); The Practical Education Project of Fuyang Normal University: Research on integrated English teaching practice in college, middle school and primary school based on "aesthetic education" (2022SJJYZX09); The Planned Project of Educational Science of Fuyang Education Bureau: Practice and exploration of middle school English teaching mode combining reading and writing from the perspective of "intertextuality" theory (FJK23028).

REFERENCES

[1] ZHU, QIUYAN.*Empowering language learning through IoT and big data: an innovative English translation approach.* Soft Computing, 2023, 27.17: 12725-12740.

[2] CHOWDHARY, K. *Natural language processing.* Fundamentals of artificial intelligence, 2020, 603-649.https://doi.org/10.1007/978-81-322-3972-7_19

[3] Diao, L.; Hu, P.*Deep learning and multimodal target recognition of complex and ambiguous words in automated English learning system.* Journal of Intelligent & Fuzzy Systems, 40(4), 2021, 7147-7158.https://doi.org/10.3233/JIFS-189543.

[4] Han, R.; Yin, Y.*Head-Driven English Syntactic Translation Model Based on Natural Language Processing.* In 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) 2021, (pp. 1244-1249). IEEE.https://doi.org/10.1109/IPEC51340.2021.9421111

[5] He, T.*ON TRANSLATION TEACHING THEORY AND TRANSLATION SKILLS IN COLLEGE ENGLISH BASED ON COGNITIVE IMPAIRMENT.* Psychiatria Danubina, 34(suppl 1), 2022, 706-708.

[6] Jingchun Zhou.; Jiaming Sun.; Weishi Zhang.; Zifan Lin.*Multi-view underwater image enhancement method via embedded fusion mechanism.* Engineering Applications of Artificial Intelligence, 2023, 121, 105946https://doi.org/10.1016/j.engappai.2023.105946.

[7] Li, B.*Research on English Translation Based on Recursive Deep Neural Network.* In 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture, 2021, (pp. 483-487).https://doi.org/10.1145/3495018.3495104

[8] Li, X.; Liu, L.; Tu, Z.; Li, G.; Shi, S.; Meng, M. Q. H.*Attending from foresight: a novel attention mechanism for neural machine translation.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2021, 2606-2616.https://doi.org/10.1109/TASLP.2021.3097939

[9] Liu, J.; Tang, B.*GDASC: Assessment of urban land use efficiency in Hebei Province based on data envelopment analysis.* International Journal of Cooperative Information Systems, 2023. https://doi.org/10.1142/S0218843023500132.

[10] Matsui, T.; Suzuki, K.; Ando, K.; Kitai, Y.; Haga, C; Masuhara, N.; Kawakubo, S.*A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders.* Sustainability Science, 17(3), 2022, 969-985.https://doi.org/10.1007/s11625-022-01093-3

[11] Őzcan, F.; Quamar, A.; Sen, J.; Lei, C*Efthymiou, V.: State of the art and open challenges in natural language interfaces to data.* In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020, (pp. 2629-2636).https://doi.org/10.1145/3318464.3383128

[12] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sutskever, I*Learning transferable visual models from natural language supervision.* In International Conference on Machine Learning, 2021, (pp. 8748-8763). PMLR

[13] Zheng, Y.*Strategies to Improve the Effectiveness of College English Translation Teaching.* Advances in Vocational and Technical Education, 3(2), 2021, 86-91.

[14] Zhu, Qiuyan.*"Empowering language learning through IoT and big data: an innovative English translation approach."* Soft Computing 27.17 (2023): 12725-12740.

[15] Wang, Yuhua. *"Artificial Intelligence technologies in college English translation teaching."* Journal of psycholinguistic research 52.5 (2023): 1525-1544.

[16] ALHALANGY, Abdalilah, et al.*Exploring the impact of AI on the EFL context: A case study of Saudi universities. Alhalangy, AGI, AbdAlgane, M.(2023).* Exploring The Impact Of AI On The EFL Context: A Case Study Of Saudi Universities. Journal of Intercultural Communication, 2023, 23.2: 41-49.

[17] MAHYOOB, Mohammad; AL-GARAADY, Jeehaan; ALBLWI, Abdulaziz.*A proposed framework for human-like language processing of ChatGPT in academic writing.* International Journal of Emerging Technologies in Learning (iJET), 2023, 18.14.

[18] BASKARA, Risang, et al.*Exploring the implications of ChatGPT for language learning in higher education.* Indonesian Journal of English Language Teaching and Applied Linguistics, 2023, 7.2: 343-358.

[19] YOO, Jiseung; KIM, Min Kyeong*Using natural language processing to analyze elementary teachers' mathematical pedagogical content knowledge in online community of practice.* Contemporary Educational Technology, 2023, 15.3: ep438.

[20] YUAN, Qiwei; DAI, Yu; LI, Guangming.*Exploration of English speech translation recognition based on the LSTM RNN algorithm.* Neural Computing and Applications, 2023, 35.36: 24961-24970.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jul 19, 2024
Accepted: Sep 2, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1363–1376, DOI 10.12694/scpe.v26i3.4354

DEEP LEARNING AND SUPPLY CHAIN BASED ENTERPRISE STRATEGIC MARKETING OPERATION MANAGEMENT SYSTEM CONSTRUCTION

XIAOTENG MA*AND LIUFENG WANG†

Abstract. The present focus of supply chain management is on how to understand customer information, establish customer segmentation, and make corporate resource allocation rational under restricted resources in order to maintain steady development. This paper explores integrating deep learning with supply chain management to enhance strategic marketing operations. It introduces an evaluation index system using a five-dimensional balanced scorecard and proposes a performance evaluation model based on the LMBP algorithm for efficient network weight calculation and training. Empirical testing with T Mobile demonstrates the model's effectiveness, achieving a 97.29% accuracy rate and over 93% empirical fit accuracy, highlighting its potential for optimizing strategic marketing operations.

Key words: Deep learning; Supply chain management; Corporate strategy; Marketing operations; Business process modeling

1. Introduction. Under the macro market background, enterprises are facing increasingly fierce competition. The comprehensive analysis of the market and the surrounding environment shows that the challenges faced by enterprises mainly come from diverse and personalized customer needs, increasingly high delivery requirements, more and more high-tech use, shorter and shorter product life cycle and globalization of market competition [1]. As a ”third party source of profit”, the role of operation management in economic activities is becoming more and more obvious, and has gradually become the most important successful competitiveness today [2, 3].

In the competitive market landscape, effective customer segmentation and resource allocation are crucial for gaining a competitive edge. Companies need to leverage data to analyze customer needs, target appropriate customer groups, and focus their supply chain resources to develop effective competitive strategies. However, resource imbalances pose significant challenges, limiting overall operational efficiency and necessitating balanced resource allocation to meet fluctuating customer demands and optimize benefits with minimal resource use.

This paper addresses these challenges by exploring how companies can enhance their strategic marketing operations through optimized resource allocation. It highlights the diversity of enterprise marketing resources, which include human, material, and financial resources, and categorizes them into internal and external resources. The paper emphasizes the importance of balancing tangible resources (such as fixed assets and cash) with intangible ones (such as brand influence and intellectual property).

A key innovation of this study is the application of contemporary marketing theories and data analysis models to improve resource allocation strategies. It proposes using advanced analytical algorithms to address critical issues in marketing information analysis, including product marketing plans, pricing strategies, and cost management. The study aims to provide a comprehensive framework for optimizing resource allocation and enhancing marketing management capabilities, particularly for companies facing rapid market changes and resource constraints.

By establishing a robust product cost information management system and integrating advanced information technology tools, this paper offers practical solutions to enhance strategic marketing operations and achieve stable development in a competitive environment.

In summary, the construction of a strategic marketing operation management system for companies based on deep learning and supply chain has important practical significance.

[*] College of Information Science and Engineering, Shandong Agricultural University, P.R. China (a8617667122022@163.com)

[†] Jiaxing Xiuhu School, Jianxing 314000, Zhejiang Province, P.R. China. (Manina202405@163.com).

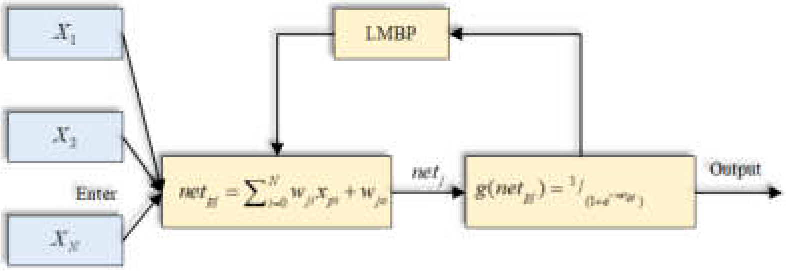

X_1
X_2
Enter
X_N
LMBP
$net_{pj} = \sum_{i=0}^{N} w_{ji} x_{pi} + w_{jo}$
net_j
$g(net_{pj}) = 1/(1+e^{-net_{pj}})$
Output

Fig. 2.1: LMBP neural network.

"This article is structured as follows: The Introduction provides an overview of the study's background and significance. The Data and Methods section outlines the study's design, participant selection, data collection procedures, and analytical methods. The Results section presents the findings of the study, focusing on the prevalence and impact of cancer-related fatigue, anxiety, and depression in hematologic cancer patients undergoing chemotherapy. Finally, the Discussion interprets the results in the context of existing literature, discusses the implications for clinical practice, and suggests potential areas for future research."

2. Mathematical model of enterprise strategic marketing operation management algorithm. The improvement and optimization of BP algorithm can be based on the weight adjustment and function change of neural network algorithm, as shown in Figure 2.1.

The sum of the mean squared errors of the network is defined as Eq2.1

$$E = \frac{1}{2} \sum_{P=1}^{N} \sum_{k=1}^{M} (y_{pk} - \hat{y}_{pk}) \tag{2.1}$$

N represents the number of input nodes, M represents the number of output units, Y represents the target output, and $\hat{y}$represents the network output. In the BP network training, the literature uses the fastest descent method and the error function minimization E in the FNN network for the multilayer network containing M hidden layers, the network of strategic marketing operation management system of enterprises is established as Eq2.2-Eq2.5.

$$net_{pj} = \sum\nolimits_{i=0}^{N} w_{ji} x_{pi} + w_{jo} \tag{2.2}$$

$$g(net_j) = 1 \Big/ _{(1+e^{-net_{pj}})} \tag{2.3}$$

$$net_{pk} = \sum\nolimits_{j=0}^{M} W_{kj} g(net_{pj}) + W_{ko} \tag{2.4}$$

$$\hat{y}_{pk} = g(net_{pk}) \tag{2.5}$$

where M is the number of hidden layers, W_{kj} is the weight value connecting node j in the hidden layer to node k in the output layer, W_{kD} is the threshold value in layer K , and $\hat{y}_{pk}$ is the target output of the k^{th} . Combining Eq2.5, Eq2.6 and Eq2.7 are derived as:

$$\Delta W_{kj} = -\eta^{\partial E} / \partial W_{kj} \tag{2.6}$$

Table 2.1: Random sample point array of enterprise strategic marketing operation management model.

Serial number	Factor 1	Factor 2
1	0.4295	0.2569
2	0.2572	0.0099
3	0.2975	0.5328
4	0.4248	0.2786
5	0.1193	0.9463
6	0.49527	0.3928
7	0.7065	0.0248
8	0.2437	0.6715
...	...	...

$$\Delta W_{kj} = -[H + \mu I]^{-1} J^T e \tag{2.7}$$

In this study, various resource allocation models and methods are reviewed, each with its strengths and weaknesses. Traditional models often rely on subjective criteria and static parameters, which can lead to inefficiencies in dynamic market conditions. For instance, while linear programming models are useful for optimizing resource allocation under fixed constraints, they may not adapt well to rapid changes in customer demand or supply chain disruptions.

Conversely, advanced methods like the LMBP algorithm used in this study offer significant advantages. The LMBP algorithm enhances training speed and accuracy in neural networks, providing more responsive and adaptive resource allocation. Its strength lies in its ability to quickly adjust network weights based on real-time data, which is crucial for managing dynamic and complex supply chain environments.

However, the LMBP algorithm is not without limitations. It requires substantial computational resources and may not perform optimally with limited data or in highly volatile conditions. Therefore, while it represents a significant advancement over traditional methods, its effectiveness is contingent on the quality and quantity of input data and the specific context of its application.

In this study, the choice of the LMBP algorithm is justified by its superior performance in empirical testing with T Mobile, demonstrating high accuracy and fit. This choice reflects the study's objective to provide a more adaptive and precise resource allocation model that can better handle the complexities of modern supply chains.

By integrating these advanced methods, this study addresses the gaps left by conventional approaches, offering a more robust framework for optimizing strategic marketing operations and resource allocation.

Then, the K-Means clustering algorithm was first applied to conduct a pre-experiment to observe the results of the algorithm operation of the strategic marketing operation management model of the enterprise [13]. 100 sample points with dimension 2 were randomly generated, as shown in Table 2.1 .

Then the program was written in MATLAB R2018a, and the number of simulations was set to 100, and the operating result graphs of each algorithm were obtained, as shown in Fig. 2.2.

It is obvious from Fig. 2.2 that each algorithm is easy to fall into the local optimal solution of the enterprise strategic marketing operation management model, and the results are unstable. In this paper, to address this defect, consider that Clara algorithm can start from K-Mediods algorithm to obtain better clustering center and have higher efficiency, here Clara algorithm is used to optimize K-Means algorithm, and the proposed improvement algorithm is named CK clustering integration algorithm, the principle is shown in Fig. 2.2 and the detailed steps of the improvement algorithm are as follows: first, the input data set c of the improved algorithm is obtained after using the normalization method, and for each attribute j , the following normalization Eq2.8 is executed.

$$C_i(j) = \frac{N_i(j) - \min(N(j))}{\max(N(j)) - \min(N(j))} \tag{2.8}$$

The centers(i) obtained in the pre-run is then used as the clustering centers of the data set, and the clustering results are output and the total error is calculated; the total error is calculated using the Euclidean

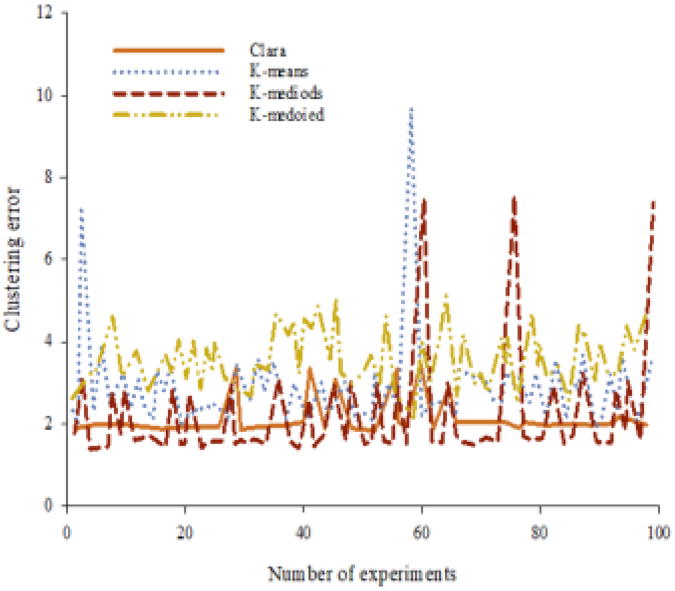


Fig. 2.2: Multiple run graphs of each algorithm.

distance, and the specific Eq2.9 is as follows.

$$dist = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{km}(C_j - c(i))^2} \tag{2.9}$$

This leads to a mathematical model and a brief procedure for the strategic marketing operations management algorithm of the enterprise, as shown in Fig. 2.3.

3. Method.

3.1. Enterprise strategic marketing resource allocation model and operation. The fundamental purpose of enterprise operation is to pursue profit maximization, and this paper is based on the principle of profit maximization to determine the total amount of marketing resources allocation. From the integration of literature on marketing resources in Chapter 2, it is clear that advertising and promotion account for most of the marketing budget and have a direct impact on sales volume when enterprises actually carry out marketing activities [14]. Promotional and advertising efforts are just as successful in the automobile industry. The primary resources for marketing efforts are financial, human, and material resources. Since dealers represent the majority of SMEs, maintaining the dealer network and growing the market also constitute major marketing costs. Furthermore, one of the biggest costs is after-sales service, which includes setting up an establishment for after-sales service stations, paying salaries to after-sales service staff, etc. All of these things call for coordinated strategic management and resource allocation for marketing inside the company.

First, we define the human resource cost H , which mainly includes the marketing staff salary cost h , travel cost b , annual training cost x , where the quantitative relationship is , and thus the Eq3.1.

$$\begin{cases} H = h + b + x \\ x = \alpha h \quad 0.1 \le \alpha \le 0.2 \end{cases} \tag{3.1}$$

Then, we define the cost of physical resources M , mainly including the cost of marketing staff office hardware equipment m_1 , office space cost m_2 , the constraint relationship for the average depreciation rate of

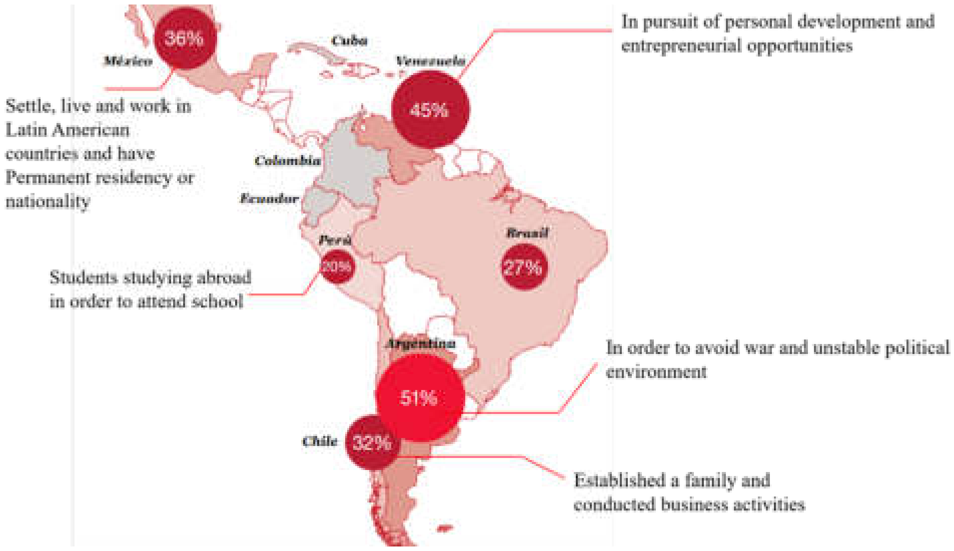

36%
México
Cuba
Venezuela
In pursuit of personal development and entrepreneurial opportunities
45%
Settle, live and work in Latin American countries and have Permanent residency or nationality
Colombia
Ecuador
Perú
Brasil
20%
27%
Students studying abroad in order to attend school
Argentina
In order to avoid war and unstable political environment
51%
Chile
32%
Established a family and conducted business activities

Fig. 2.3: Mathematical model and brief procedure of strategic marketing operations management algorithm for enterprises.

hardware equipment $0.1 < \beta < 0.2$, with Eq3.2.

$$\begin{cases} M = \beta m_1 + m_2 \\ 0.1 \leq \beta \leq 0.2 \end{cases} \tag{3.2}$$

Define the cost of financial resources F , mainly including advertising costs A , market development costs G , office supplies costs d , and business hospitality costs e . Eq3.3 is obtained.

$$\begin{cases} J = j_1 + j_2 \\ J_1 = \theta Q \\ 0.005 \leq \theta \leq 0.01 \end{cases} \tag{3.3}$$

The strategic marketing operations management cost function of the firm under the model was finally determined as Eq3.4.

$$C = f(H, M, F, J, S) = H + M + F + J + S \tag{3.4}$$

Then according to the previous description of the enterprise strategic marketing operation management model, this paper proposes a five-dimensional balanced scorecard-based enterprise strategic marketing operation management and supply chain evaluation model, and its simplified process is shown in Fig. 3.1.

For the model mentioned in this study, its establishment and optimization can be systematically summarized into three stages: data processing, neural network construction and subsequent processing application. The analysis steps are shown in Fig. 3.2.

3.2. Empirical model architecture. In the methodology of this study, Logistic Regression Analysis and K-means Cluster Analysis are employed sequentially, each serving a distinct purpose that contributes to the overall analysis.

Logistic Regression Analysis. This step is utilized first to identify the key predictors or feature variables that significantly influence the outcomes of interest, such as customer purchasing decisions or product demand. By determining the relationships between these variables and the likelihood of certain behaviors (e.g., purchase likelihood), Logistic Regression helps to refine the dataset by highlighting the most relevant factors. This step is critical as it informs the selection of variables that will be used in the subsequent clustering process, ensuring that only the most impactful data is considered.

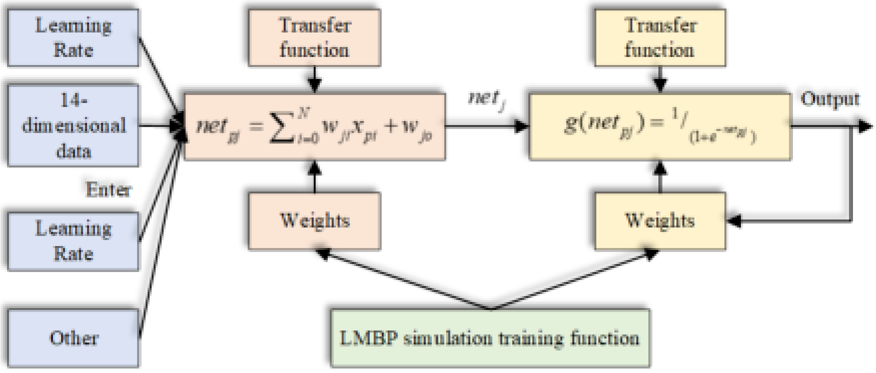

Learning Rate
14-dimensional data
Enter
Learning Rate
Other
Transfer function
$net_{pj} = \sum_{i=0}^{N} w_{ji}x_{pi} + w_{jo}$
Weights
net_j
Transfer function
$g(net_{pj}) = 1/(1+e^{-net_{pj}})$
Output
Weights
LMBP simulation training function

Fig. 3.1: Neural network for strategic marketing operation management of enterprises based on supply chain performance evaluation.

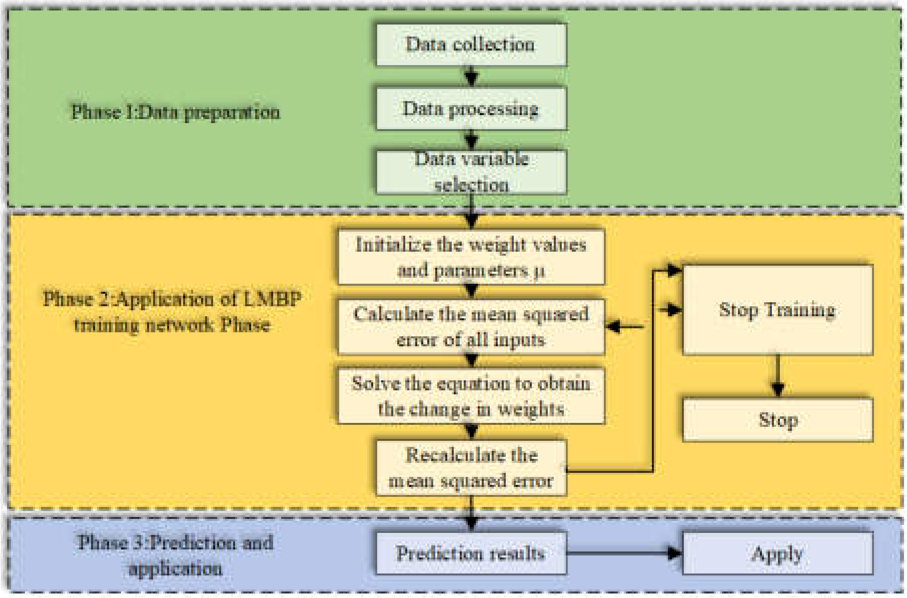

Phase 1:Data preparation
Data collection
Data processing
Data variable selection
Phase 2:Application of LMBP training network Phase
Initialize the weight values and parameters μ
Calculate the mean squared error of all inputs
Solve the equation to obtain the change in weights
Recalculate the mean squared error
Stop Training
Stop
Phase 3:Prediction and application
Prediction results
Apply

Fig. 3.2: The process of establishing the strategic marketing operation management model of the enterprise.

K-means Cluster Analysis. Following Logistic Regression, K-means Cluster Analysis is applied to the refined dataset. This clustering technique groups similar entities—such as customers or products—based on the significant feature variables identified in the previous step. The purpose of this analysis is to segment the dataset into distinct clusters, each representing a unique group with similar characteristics. This segmentation allows for more targeted marketing strategies and resource allocation, as the clusters represent groups with shared behaviors or needs.

Relationship and Sequence. The sequence of these steps is crucial for effective analysis. Logistic Regression Analysis first identifies which variables are most influential, thereby reducing noise and focusing the dataset on the most relevant information. K-means Cluster Analysis then uses this focused dataset to create meaningful clusters. The effectiveness of the clustering process is enhanced by the prior logistic regression, as it ensures that the clusters are based on variables that truly matter, leading to more actionable insights.

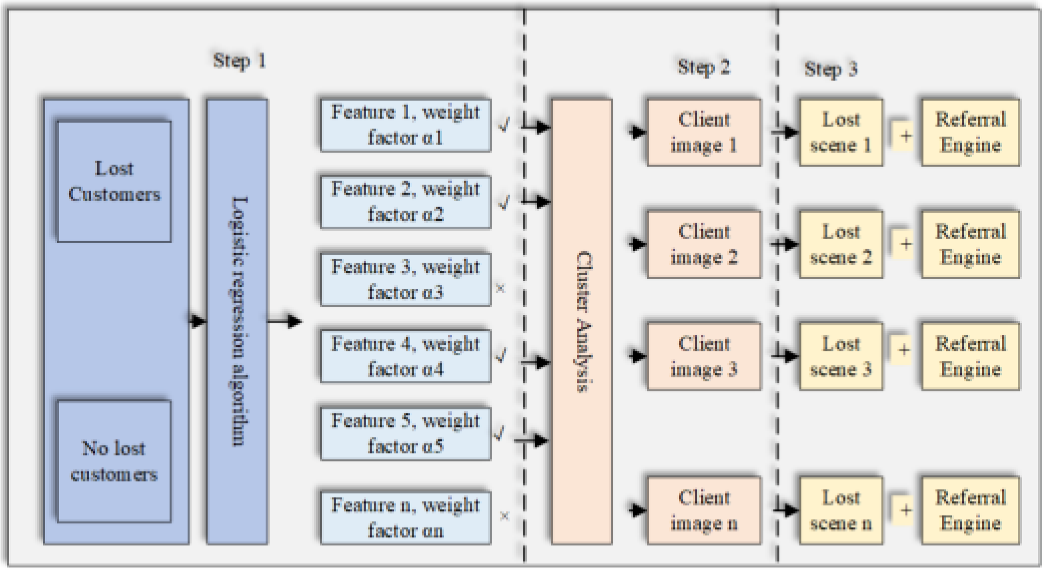


Fig. 3.3: Application ideas of strategic marketing operation management model for enterprises.

By clearly outlining the purpose of each step and their order of precedence, the study ensures that the methodological approach is transparent and logically structured, ultimately enhancing the validity and utility of the findings.

In the enterprise strategic marketing operation management model, the construction is divided into Logistic regression analysis step and K-means algorithm customer portrait clustering analysis step, in which Logistic regression analysis step, the main idea is to build a data mining model by historical lost customers, and use the feature importance analysis ability of Logistic regression algorithm to output each feature variable in In the K-means clustering algorithm stage, after the features that have an important impact on customer churn are selected in the previous step, the clustering analysis is performed on the churned customers based on these feature variables, and the clustering results can be used to cluster potential churned customers [15]. The clustering results can further guide the construction of the churn maintenance operation system. The idea of applying the strategic marketing operation management model is shown in Fig. 3.3.

We have selected T Mobile to be the empirical test object for this study's empirical test phase. This paper uses T Mobile's telephone outbound marketing data from September 2020 to December 2020 in order to ensure that the findings can help with T Mobile's marketing management design, strategic positioning, and customer churn, as well as to ensure the validity and truth of the research data content. However, due to the large number of customers, 228,778 households were randomly selected from them without affecting the model effect marketing data for the study. In order to ensure the accuracy and generalization ability of the model prediction, this study will cover all kinds of behavioral characteristics of users and information characteristics of marketing services as much as possible, and obtain data including users' basic information such as network length, gender and age, business information such as Internet traffic, call length and active days, tariff information such as tariff level and contract bundle, recommended strategy level, recommended strategy traffic expansion capacity and other marketing information. Marketing information, finally involving a total of 125 variables.

The following are the annotations of some fields of the data, such as the annotation of the basic user information field in Table 3.1, which contains information on user's gender, age, length of time in the network, etc. Table 3.2 is the annotation of the user contract information field, which records the subscription characteristics of the user contract bundled products; Table 3.3 is the annotation of the user business behavior information field, which records the information related to the use and activity of the user's Internet behavior and calling behavior; Table 3.4 is the annotation of the subscription information field, which records the information related

Table 3.1: User basic information field comments.

Fields	Note
Serv _ number	Cell phone number
Gender _id	Gender
Age	Age
Join _duration	Length of time in the network (months)
Credit _class _id	Star level
Wb _bind _user_kind	Whether to converge users
Cmasp _asp _type	DM single mobile single extranet extranet operator
Number _ame _id	Number of numbers under the same ID card
User_owned _branch	User belongs to the branch

Table 3.2: User Contract Information Field Comments.

Fields	Note
Serv _ number_ hy	Contract user or not
Contract _ month _ owe	Remaining duration of the contract

Table 3.3: User business behavior information field comments.

Fields	Note
Minimum _ contract _ spending	Minimum monthly consumption of terminal contract

Table 3.4: User tariff subscription information field comments.

Fields	Note
Inner _ gprs _ pct	Current Month Traffic Saturation
Inner _ gsm _ pct	Voice saturation for the month
M _ gprs _ bill _ flux _ dtal	Current month billed traffic usage
Avg _ jf_ dou _ 3mon	Average monthly billed traffic usage in the past three months
R _ jf _ dou _ 3mon	Monthly average billed traffic usage volatility in the past three months
Dou	Current Month DOU
Dir _ dou	Monthly targeted traffic usage
Cur _ mon _ voice _ days	Number of call days in the month
If _ flux _ ct	Whether the current month traffic over-set users
If _ gsm _ ct	Whether the current month is a voice over-set user
If _ off _ flux _ ct	Whether it is a frequent traffic over-set user
If_ off _ gsm _ ct	Whether it is a frequent traffic voice user
Cur _ mon _ flux _ days	Number of days of traffic use in the month

to the user's current subscription tariff service and the content of the tariff service; Table 3.5 is the annotation of the customer marketing information field, which records the service information of the recommended strategy when actively marketing to the customer and the service expansion information relative to the user's current subscription tariff, etc.

4. Case study.

4.1. Model optimization validation. The data set of retail customer orders is processed using Python, and the final segmentation variables for 36763 customers are shown in Fig. 4.1, Fig. 4.2, and Fig. 4.3.

The algorithm program is written in MATLAB R2018a, and finally the algorithm performance is verified

Table 3.5: Customer Marketing Information Field Comments.

Fields	Note
Mkt_user_group_desc	Operation Strategy
If_success_deal	Whether the operation is successful
Discnt_type	Strategy offer combination type
Strategy_type	Strategy Type
Inclu_flu	Strategy information_contains preferential traffic
Inclu_dur	Strategy information_contains preferential voice
Inclu_dir_flu	Strategy information_includes preferential targeted traffic
Fee_front Policy Information_Phase 1	Price After Discount
Fee_after Policy Info_Phase 2	Discounted Price
Period front Policy Info_Phase 1	Discount Length
Period after Policy_Phase 2	Discounted hours

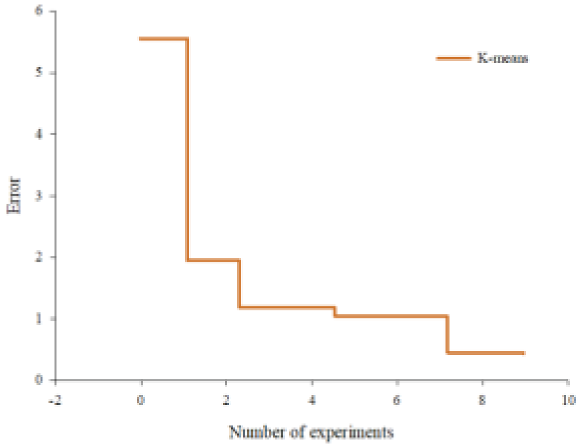


Fig. 4.1: Iteration curve of K-Means algorithm.

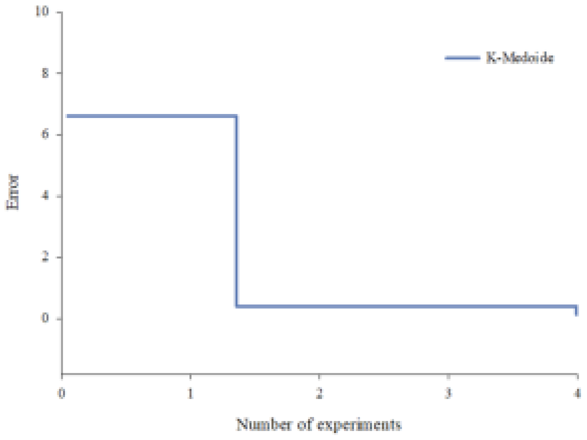


Fig. 4.2: Iteration curve of K-Medoide algorithm.

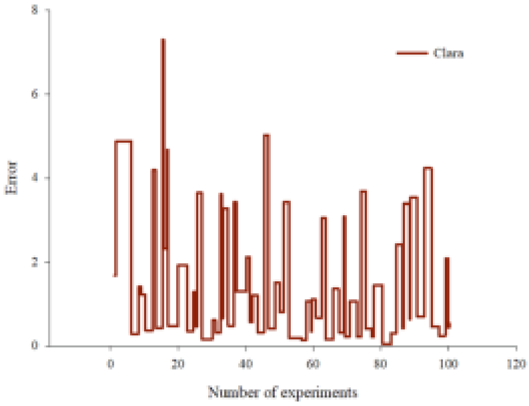


Fig. 4.3: Clara algorithm iteration curve.

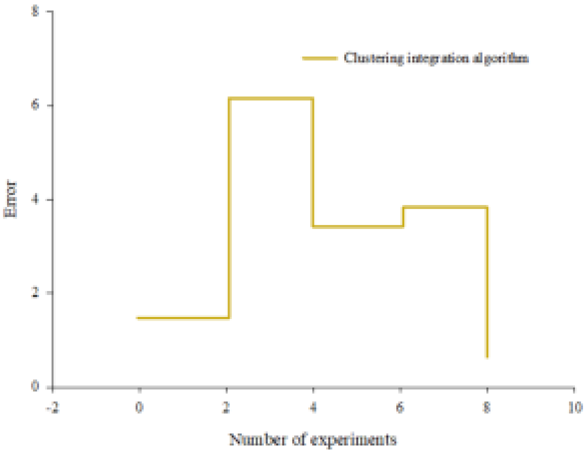


Fig. 4.4: Comparison of total error before and after algorithm optimization.

by using algorithm comparison in the case of algorithm validation. The running iteration curve obtained by using the customer clustering factors as the input data of the algorithm is shown in Fig. 4.4, which shows that the improved CK clustering integration completes the iteration in the 8th round, and the iteration speed is more blocky when compared with K-Means algorithm, K-Medoide algorithm, K-Mediods algorithm, and Clara algorithm. To verify the performance of the CK clustering integration algorithm with the remaining algorithms, the optimization resulted in a significant 13% reduction in system error, as shown in Fig. 4.4.

Based on the prediction results of the model on the test set, the model prediction effect evaluation index is obtained.

The check accuracy rate is:

$$\Pr ecision = \frac{TP}{TP + FP} = 97.29\% \tag{4.1}$$

According to the prediction results, it can be seen that the logistic regression model established for churn customer prediction has a good prediction effect, so the weight coefficients of each feature variable output from the model can be used as a basis for judging the importance of the features. Based on the results of the run, we generally consider the features with P-values less than 0.05 to be significant for model prediction, and Table

Table 4.1: Results of the program run to calculate the weight coefficients of the characteristic variables.

if_ct_cur_day	Whether the current month is over set	-1.444621180	0.603648488	-2.393148752	0.0016803322
if_jmb_cur_day	Is there a reduction package in the current month	0.887794131	0.290237359	3.058854421	0.002312845
if_jmb_end_cur_mon	Whether the current month reduction package expires	-1.354896755	0.403436820	-3.358378352	0.000782988
if_order_wb	Whether to subscribe to broadband	1.790728519	0.794705835	2.253341623	0.025227831
if_bensheng_user	Is it a user in this province	-0.202891716	0.094718683	-2.1420544607	0.032279722
is_tj	Whether the last month is closed	-1.624844361	0.287323465	-5.649205523	0.000000015
is_wb_bind_user	Is the broadband bundle users	2.327311892	0.729575326	3.189952778	0.001422855
is_wb_bind_user_lst1	Whether broadband users last month	3.297398250	0.132989531	2.910351629	0.003510239
is_wb_bind_user _Ist2	Whether the last month is broadband users	-3.4207775	1.089823560	-3.167923325	0.001525373
is_pre_prd_chng_down	Is the package downgrade users	2.675488950	1.222509216	2.188533576	0.027521566
effect_date	Length of time on the network	0.281032752	0.107003748	-3.552268892	0.008520032
last_voice_to_now	Duration of last call to current	-0.152847928	0.043026947	2.8869788769	0.000381687
cur_mon_voice_days	Number of days of voice communication	0.439566140	0.1522583219	2.9732159823	0.003789592
slnt_days	Number of days of silence in the current month	0.17419289	0.058577137	2.973690351	0.002932427

4.1 shows all the variables with P-values less than 0.05 among the results obtained from this program run.

In the above results, Estimate is the estimated value of the importance coefficient of each characteristic variable, Std.Error is the standard error of the coefficient estimate, z value is the z statistic, $\Pr(> |z|)$ is the estimated p value of the characteristic variable, and the smaller p value means the more important to the outcome variable, it can be seen that the model proposed in this study has a better fit both in terms of accuracy optimization and empirical testing.

4.2. Analysis of empirical results. Based on the model calculated in Chapter 4 on the training sample set, the ROC plot shown in Fig. 4.5 is obtained after validation on the test sample set.

As can be seen from the figure, the AUC value of the model calculated based on the above parameters is 80.1%. In a general classifier, the ideal classifier does not produce any prediction errors, i.e., the model can achieve a 100% true positive rate before producing any false positives, at which point the AUC value is 1. In a random classifier, each correct prediction is followed by an incorrect prediction the next time, at which point the AUC value is 0.5. So, however, from the ROC curve, a model with an AUC > 0.8 classifies effect is quite good. Another way to determine whether the final trained model is reasonable is to observe the importance of each feature variable in the XGBoost model calculation, and the top 10 important features according to the function calculation are listed in Table 4.2.

From Table 4.3, it can be seen that whether the user for marketing recommendation strategy, mainly affected by the following aspects: the level of preferential recommendation strategy. In general, the sales price of the recommended product in a marketing recommendation is essentially the same as the user's current stable consumption level; that is, the higher the recommended strategy traffic expansion capacity, the more affordable it is for the user. These two characteristics of importance, the recommended strategy traffic expansion capacity and the type of discount reduction, both represent the level of benefits to the user marketing recommendation strategy. as well as the kind of discount decrease, which describes the union of tariff strategy, discount duration, and The tariff strategy combination's kind of preference length is referred to as the type of preference reduction, includes 12 months, 24 months, "3+9", "6+6", "12+12" and other types of preference length combinations, and

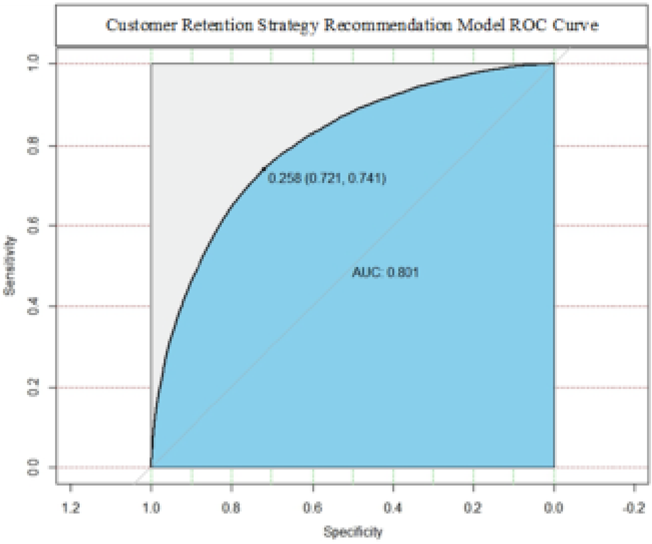


Fig. 4.5: ROC curve of the recommended model of customer retention strategy based on strategic marketing operation management of the company.

Table 4.2: XGBoost Output Feature Importance.

Feature Name	Chinese Comments	Feature Importance
Inclu_flh_up	Recommended strategy traffic expansion capacity	0.0469
Period_type	Preference reduction type	0.0422
If_wary	Whether King's Glory Preferred User	0.0340
Inner_gsm_pct	In-suite voice saturation	0.0333
Lasti1_dou	Last month's Internet traffic	0.0330
M_gprs_bil_flux_dtal	The amount of billed traffic usage for the month	0.0309
Credit_class_id	Star Rating	0.0309
If_contract_terminal	Whether the current terminal is a contract terminal	0.0267
Join_duration	Length of time on the network	0.0276
Incr_data_gprs_fee	Data traffic charge for the month	0.0230

Table 4.3: Comparison of prediction effects of different algorithms.

Models	TP	TN	EP	FN
CART	4789	2907	8611	3247
Logistic Regression	4362	2892	9035	2852
Random Forest	5165	2912	8239	3126
XGBOOST	1225	3183	2169	2638

users' concern about the type of preference reduction also It also shows that the level of recommendation strategy offer plays an important role in the success of marketing recommendation.

Based on the above ideas, we subjected the test set to model validation by running three prediction models, CART decision tree, logistic regression, and random forest, respectively, and obtained the confusion matrix metrics for each type of model, and the results are shown in Table 4.3.

A bar graph based on the above data is shown in Fig. 4.6.

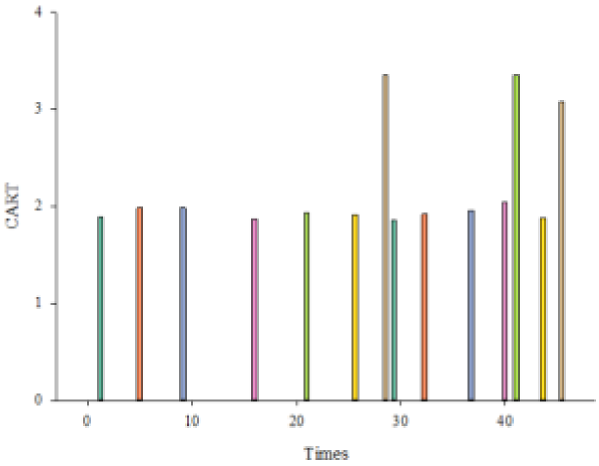


Fig. 4.6: Comparison of prediction effectiveness of different algorithms.

Table 4.4: Comparison of evaluation indexes of different algorithms.

Models / Evaluation Metrics	Accuracy Rate/%	Accuracy/%	Completeness/%	F1-score/%
CART	74.11	35.69	59.55	44.68
Logistic Regression	74.08	32.58	60.38	42.29
Random Forest	75.23	38.49	62.37	47.57
XGBOOST	93.79	83.68	94.55	88.81

Of course, it is rather thin to compare the good or bad prediction effect of various algorithms only by the value of confusion matrix, so four secondary indicators are also proposed on the separate mathematical addition method, and the results are shown in Table 4.4.

It can be seen that the accuracy of this model reaches more than 93%. By analyzing the recommendation probability of the recommendation strategy relative to the user, we can classify the strategy recommendation probability into four different levels, less than 0.2, 0.2-0.5, 0.5-0.8, and more than 0.8. When the probability of recommending a policy to a user is less than 0.2, it means that the user does not prefer the recommended policy and may have a high probability of rejecting it. When the recommendation probability of the user recommendation strategy is 0.2-0.5, it means that the user has a relative preference for the recommendation strategy, but further marketing recommendation is needed before the user can handle it. When the recommendation probability of the user recommendation strategy is 0.5-0.8, it means that the user has a high preference for this recommendation strategy, and the operator may recommend it successfully with a simple recommendation. When the recommendation strategy to the user is above 0.8, it means that the user has a very high preference for the strategy, and at this time, the high-cost outbound marketing channel can be considered to be converted into a low-cost SMS distribution channel for marketing, as expected from the empirical results.

5. Conclusion. This paper first analyzes the basic situation of enterprise strategic marketing operation management system based on deep learning and supply chain, and takes T mobile company as an empirical model, extracts historical customer marketing data of stock customers, builds a recommendation model for customer strategy preference by using classification algorithms such as XGBoost, and then analyzes the portrait characteristics of historical lost customers of T mobile company based on K-means clustering algorithm. The experimental results prove that the check accuracy of the enterprise strategic marketing operation management optimization model proposed in this study is 97.29%, the empirical fitting accuracy is higher than 93%.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

Funding Statement. There is no specific funding to support this research.

REFERENCES

[1] Kumar, T. S.*Data mining based marketing decision support system using hybrid machine learning algorithm.* Journal of Artificial Intelligence, (2020) 2(03), 185-193.

[2] Abbas, K., Afaq, M., Ahmed Khan, T., & Song, W. C.*A blockchain and machine learning-based drug supply chain management and recommendation system for smart pharmaceutical industry.* Electronics,(2020). 9(5), 852.

[3] Kamble, S. S., Gunasekaran, A., Kumar, V., Belhadi, A., & Foropon, C. *A machine learning based approach for predicting blockchain adoption in supply Chain.* Technological Forecasting and Social Change, (2021).163, 120465.

[4] Perevozova, I., Horal, L., Mokhnenko, A., Hrechanyk, N., Ustenko, A., Malynka, O., & Mykhailyshyn, L.*Integration of the supply chain management and development of the marketing system.* International Journal of Supply Chain Management,(2020). 9(3), 496-507.

[5] Valaskova, K., Ward, P., & Svabova, L.*Deep learning-assisted smart process planning, cognitive automation, and industrial big data analytics in sustainable cyber-physical production systems.* Journal of Self-Governance and Management Economics,(2021). 9(2), 9-20.

[6] Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., ... & Machtynger, L.*Artificial intelligence (AI) in strategic marketing decision-making: a research agenda.* The Bottom Line,(2020). 33(2), 183-200.

[7] Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., & McFarlane, D.*Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing.* International Journal of Production Research, 58(11), 3330-3341.

[8] Helo, P., & Hao, Y.*Artificial intelligence in operations management and supply chain management: An exploratory case study.* Production Planning & Control, (2022). 33(16), 1573-1590.

[9] SINGH, Arpit, et al.*Identifying issues in adoption of AI practices in construction supply chains: towards managing sustainability.* Operations Management Research, 2023, 16.4: 1667-1683.

[10] Huang, M. H., & Rust, R. T. *A strategic framework for artificial intelligence in marketing.* Journal of the Academy of Marketing Science,(2021). 49(1), 30-50.

[11] Hair Jr, J. F., & Sarstedt, M.*Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing.* Journal of Marketing Theory and Practice,(2021). 29(1), 65-77.

[12] Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M. C., & Johnson, J. E.*Can deep learning predict risky retail investors, A case study in financial risk behavior forecasting.* European Journal of Operational Research, (2020). 283(1), 217-234.

[13] Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., & Fujita, H.*Adaptive stock trading strategies with deep reinforcement learning methods.* Information Sciences, (2020).538, 142-158.

[14] LIU, Bochao.*Integration of novel uncertainty model construction of green supply chain management for small and medium-sized enterprises using artificial intelligence.* Optik, 2023, 273: 170411.

[15] HOSSEINNIA SHAVAKI, Fahimeh; EBRAHIMI GHAHNAVIEH, Ali.*Applications of deep learning into supply chain management: a systematic literature review and a framework for future research.* Artificial Intelligence Review, 2023, 56.5: 4447-4489.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Jul 24, 2024
Accepted: Sep 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1377–1385, DOI 10.12694/scpe.v26i3.3900

DESIGN OF ELECTRICAL LOAD PREDICTION SYSTEM BASED ON DEEP LEARNING ALGORITHM

YUANLI XU[*]

Abstract. In order to solve the problems of low prediction accuracy and long model time in current prediction algorithms, the author proposes the design of an electrical load prediction system based on deep learning algorithms. The author proposes an improved deep learning short-term load forecasting model based on random forest algorithm and rough set theory. The model is first based on historical data and uses the random forest algorithm to extract key feature quantities that affect load forecasting. Then, the key feature quantities and historical load values are used as input and output terms for training the deep neural network, and the prediction results are corrected through rough set theory. Finally, simulation verification is conducted through numerical examples. The experimental results showed that compared with the RF-DL model, the MSE index of the RFDL-RST model decreased by 30.187%, and the overall prediction results were closer to the true values. The MAE index also decreased from 5.76% to 4.02%. During special periods of significant load changes such as 07:00-08:00 (rapid increase in load) and 22:00-23:00 (rapid decrease in load), the prediction accuracy was greatly improved. In addition, compared with the DL-RST model, the MAE and MSE indicators of the RF-DL-RST model were reduced by 15.210% and 21.414%, respectively, and the DL training time of the RF-DL-RST model was shortened by 10.175%, indicating that simplifying the DL input feature quantity through the RF model can improve the load forecasting effect. The prediction accuracy of this model is higher than that of a single deep learning model and a model without prediction correction.

Key words: Electrical load forecasting, Random Forest (RF) algorithm, Deep learning (DL), Rough Set Theory (RST)

1. Introduction.

Predicting electricity demand in the near future is crucial for maintaining the safety and efficiency of power systems[1]. In recent years, with the continuous development of the global electrical market, the spot market and intraday trading system have been continuously improved, and the requirements for load forecasting accuracy have become increasingly high [2]. There are various factors that affect load, including weather factors (temperature, humidity, sunlight intensity, etc.) and time factors (working days, holidays, current specific time, etc.). At the same time, some policy factors can also lead to changes in load patterns, such as factory production reduction and shutdown caused by epidemic control, resulting in a decrease in electricity load; Encouraging policies for electric vehicles have led to an increase in electricity demand. The above factors make short-term load forecasting exhibit strong non-linear and stochastic characteristics [3]. Electrical load forecasting is a series of forecasting work that takes electrical loads as objects, including predicting future electrical demand (power), predicting future electricity consumption (energy), and predicting load curves [4]. Forecasting the load on electrical systems plays a vital role in planning and operating power grids, forming the basis for tasks like dispatching, regulation, and control. Typically, load forecasting spans various timeframes, including long-term, medium-term, short-term, and ultra short-term, each serving specific purposes within the electrical system [5,6]. Short-term load forecasting, extensively studied and highly relevant to experts and scholars, focuses on predicting electrical demand for the upcoming day to week. Typically, this involves forecasting the capacity or daily and weekly consumption data of a specific region. The aim is to establish power generation plans and inform operational scheduling. Various methodologies exist for short-term load forecasting, with intelligent forecasting methods currently being the predominant approach. This method applies powerful intelligent algorithms to establish a prediction model and make predictions. As the power grid continues to expand, its complexity grows, demanding more sophisticated load forecasting techniques. Intelligent algorithms like neural networks and support vector machines are commonly employed for this purpose. As the grid becomes smarter, there's a greater need for load forecasting methods that can meet the evolving demands of this dynamic environment. Traditional intelligent algorithms belong to shallow structure algorithms, and shallow structures

[*]Weifang Vocational College, Weifang, Shandong, 262700, China (Corresponding author, wzyjdxu@163.com)

are difficult to effectively represent nonlinear complex functions when given limited samples [7].

2. Literature Review. With the rapid progress of computer technology, machine learning is undergoing a resurgence, finding applications across diverse fields like image recognition, object detection, and natural language processing. In electrical load forecasting, machine learning algorithms have demonstrated considerable success. Advanced techniques such as reinforcement learning and transfer learning have been leveraged in this domain. However, traditional machine learning approaches face challenges with high-dimensional data. Deep learning, particularly Artificial Neural Networks (ANNs), addresses this by extracting meaningful features from complex data, thereby enhancing forecasting accuracy. ANNs, consisting of input, output, and hidden layers, serve as foundational tools in machine learning, facilitating the modeling of intricate relationships within the data. Sasidharan, M. P. et al. conducted an analysis of several prominent deep learning architectures, including Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), as well as hybrid models combining LSTM with CNN, and CNN with GRU, among others. They explored the suitability of these models for predicting charging load in charging stations. Training data were sourced from multiple charging stations within a specific region to train the models. The performance of each model was evaluated using standard metrics, and the resulting predictions were thoroughly assessed and presented [8]. Jin, X. B. et al. introduced an attention-based encoder-decoder network enhanced by Bayesian optimization to achieve precise short-term power load forecasting. This model adopts an encoder-decoder structure incorporating a Gated Recurrent Unit (GRU) recurrent neural network, renowned for its robustness in handling time series data [9]. Zhu, C. et al. introduced a power load forecasting approach utilizing Long Short-Term Memory (LSTM) networks. Through comparative analysis with conventional models, their method demonstrated reduced errors and increased applicability, showcasing its effectiveness in power load prediction [10].

The author proposes an improved deep learning (DL) short-term load forecasting model (RF-DL-RST) based on RF algorithm and rough set theory (RST). This model introduces policy factors and, together with time and weather factors, establishes a load forecasting feature set. Key feature quantities and historical load values are used as inputs and outputs for deep learning training, and the prediction results are corrected through rough set theory. Simulate and verify the effectiveness of the model.

3. Method.

3.1. Introduction to Random Forest Algorithm. The schematic diagram of the random forest algorithm is shown in Figure 1. The key to the random forest algorithm lies in the decision tree, which obtains prediction or regression results by voting or weighted averaging the prediction results of each decision tree [11].

Researchers both domestically and internationally have developed numerous decision tree algorithms, including ID3, C4.5, and Classification and Regression Tree (CART). These algorithms employ a top-down methodology to construct decision trees [12]. In the process of forming a decision tree, each new node needs to choose a new attribute as the basis for splitting. The difference between these three decision tree algorithms lies in the decision criteria for leaf splitting during the growth process. Among them, CART uses minimum mean square error as the attribute metric for splitting regression trees and Gini index (GI) as the splitting criterion for classification trees. When using the random forest algorithm for classification, the final result is determined by voting. When using the random forest algorithm for regression, the prediction result is obtained by taking the mean [13,14]. In addition, in order to reduce the impact of overfitting and random errors on the prediction results, the original data is generally divided into training and testing sets, and then the bootstrap method is used to extract the training set. Then, the CART algorithm is used to train each decision tree from top to bottom one by one until it meets the requirements.

3.2. Feature extraction. How to select key feature quantities in the dataset is crucial for reducing model complexity and shortening computation time. When extracting key feature quantities using the random forest algorithm, the Gini index or out of bag data error rate is generally used for evaluation.

The author used the Gini index method for research, and the principle is as follows: Assuming the dataset

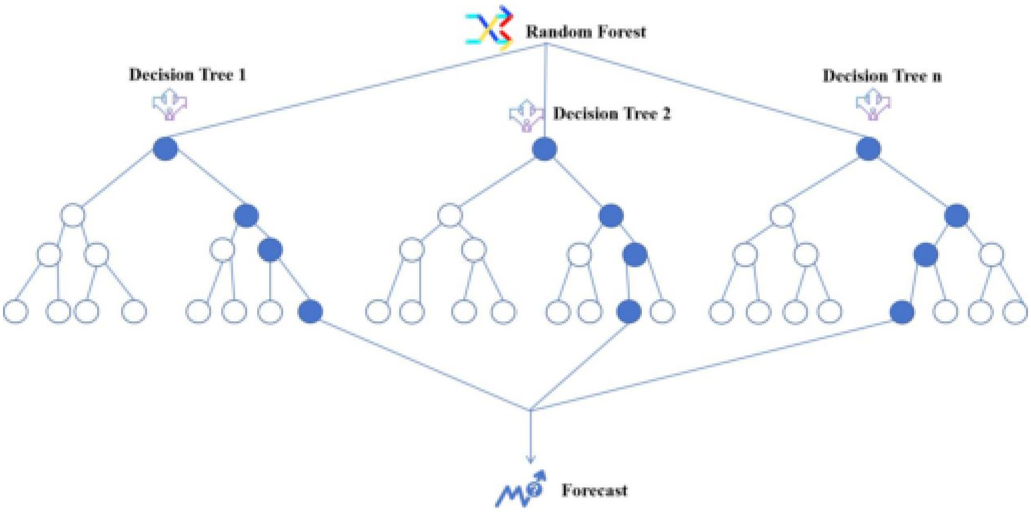


Fig. 3.1: Schematic diagram of the random forest algorithm

has J feature quantities $(X_1, X_2, X_3, \cdots, X_j)$, C categories, and I decision trees, the Gini index of node m is

$$G_m = \sum_{c=1}^{C} \widehat{p_{mc}}(1 - \widehat{p_{mc}}) \tag{3.1}$$

In the formula, $\widehat{p_{mc}}$ represents the probability estimate that node m sample is of class c.

The importance score V_{jm}^{GI} of feature quantity X_j at node m is represented by the change in Gini index before and after node m branching:

$$V_{jm}^{GI} = G_m - G_l - G_r \tag{3.2}$$

In the formula, G_l and G_r are the Gini indices of two new nodes 1 and r after node m branches.

If the set of nodes where the feature quantity X_j appears in the i-th tree is set to M, then the importance of the feature quantity X_j in the i-th tree is represented as

$$V_{ij}^{GI} = \sum_{m \in M} V_{jm}^{GI} \tag{3.3}$$

In summary, the importance of the feature quantity X_j in RF can be expressed as

$$V_j^{GI} = \frac{1}{I} \sum_{i=1}^{I} V_{ij}^{GI} \tag{3.4}$$

From this, it is possible to rank the importance of each feature quantity in the dataset and extract important feature quantities.

3.3. Principles of Deep Learning. A Deep Neural Network (DNN) is a type of neural network architecture that consists of multiple layers, including at least one hidden layer [15]. Compared with traditional BP neural networks, the two have similar structures, but DNN generally has more hidden layers and adopts a layer wise training mechanism to overcome the gradient diffusion problem in BP neural network training. Compared with traditional solving methods, well trained DNNs have higher computational efficiency and accuracy.

A typical DNN network structure consists of input and output layers at the beginning and end, with the middle layer being the hidden layer and the layers being fully connected (any node in the previous layer must be connected to any node in the following layer). Assuming there are g nodes in the i-1st layer, the output h_j^i of the jth node in the i-th layer is represented as

$$h_j^i = \sigma(z_j^i) = \sigma(\sum_{k=1}^{g} \omega_{jk}^i h_k^{i-1} + b_j^i) \tag{3.5}$$

In the formula: $\sigma(\cdot)$ is an activation function used to sum and further enhance the input of a node; ω_{jk}^i is the weight coefficient from the k-th node in layer i-1 to the j-th node in layer i; h_k^{i-1} is the output of the k-th node in the i-1 layer; b_j^i is the deviation coefficient of the jth node in the i-th layer.

The author uses the mean squared error loss function, represented as follows:

$$L = \frac{1}{PT} \sum_{p=1}^{T} \sum_{t=1}^{T} (y_{p,t} - p'_{pt})^2 \tag{3.6}$$

In the formula: P represents the number of training samples; $y_{p,t}$ is the expected value of the p-sample at time t; $y'_{p,t}$ is the predicted value output by DNN; T is the number of predicted time periods.

At the same time, the author introduces L2 regularization to the loss function, aiming to limit the weight parameters to a certain range to adapt to outliers and noise. The expression is as follows:

$$L = \frac{1}{PT} \sum_{p=1}^{p} \sum_{t=1}^{T} (y_{p,t} - y'_{p,t})^2 + \frac{\alpha}{2} \omega^T \omega \tag{3.7}$$

In the formula: α to regularize hyperparameters; ω for weight vectors.

Set the learning rate of the parameter as μ, update the hidden layer parameters repeatedly through equation 3.7 until the prediction accuracy converges.

3.4. Predictive Correction Model. Rough set theory is a mathematical tool for dealing with uncertainty and fuzzy problems, which can effectively correct and analyze defect information that is inconsistent, requires error correction, or has data loss [16].

Establishing a load forecasting correction model using rough set theory:

$$\begin{cases} y'_{t+1} = y_{t+1} + s_t |k_{t+1} - k_t| \\ k_{t+1} = y_{t+2} - y_{t+1} \\ k_t = y_{t+1} - y_t \end{cases} \tag{3.8}$$

In the formula, y_{t+1} and y'_{t+1} are the predicted and corrected values at time t+1, respectively; s_t is the scale factor.

In order to solve the scaling factor s_t, an information system needs to be constructed. The author assumes that the information system on which the rough set theory is based is K=(U,A), where: the domain U is the set of predicted values output by DNN; $A = C \cup S$ is the attribute set, S=st represents the decision attribute, and the conditional attribute C is the set of feature quantities in the dataset. Based on existing research results, C={a,b,c} is defined here [17]. Among them:

$$a = \frac{|k_{t+1} - k_t|}{y_t} \tag{3.9}$$

$$b = sgn(k_{t+1} - k_t) \tag{3.10}$$

$$c = |\frac{y_t}{max(y_t)}| \tag{3.11}$$

At this point, the load prediction value can be corrected through equations 3.9-3.12.

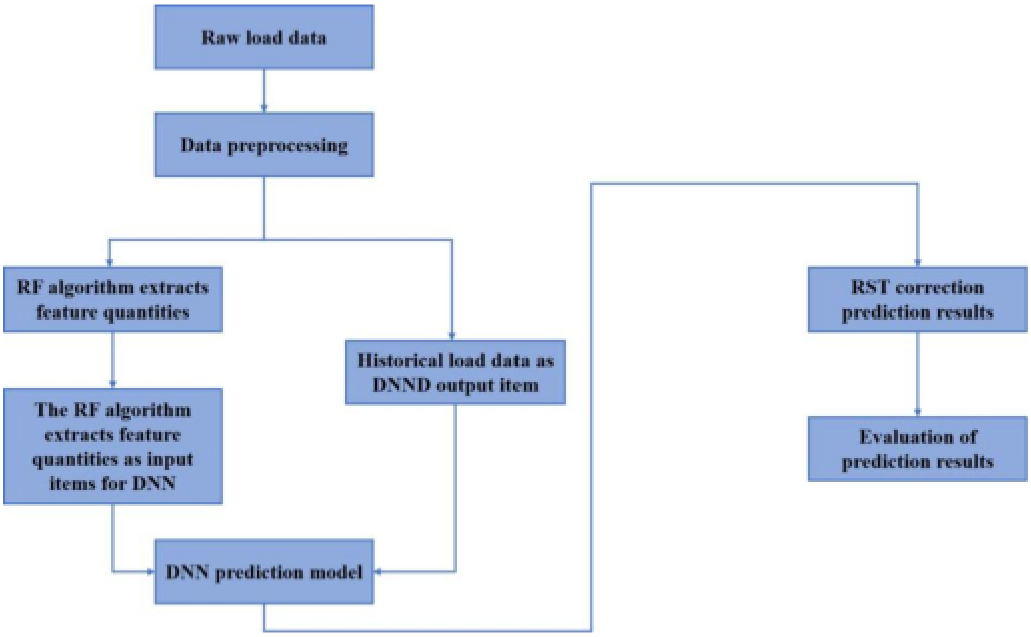

Raw load data
Data preprocessing
RF algorithm extracts feature quantities
Historical load data as DNND output item
The RF algorithm extracts feature quantities as input items for DNN
DNN prediction model
RST correction prediction results
Evaluation of prediction results

Fig. 3.2: Schematic diagram of RF-DL-RST model

3.5. Prediction Result Evaluation Model. The author employs two metrics, mean square error (MSE) and maximum absolute error (MAE), to assess the accuracy of the prediction results. MSE quantifies the overall prediction performance by measuring the average squared difference between predicted and actual loads. On the other hand, MAE evaluates the predictive accuracy at specific points by calculating the maximum absolute difference between predicted and actual loads. The MSE and MAE are indicated as follows:

$$\epsilon_{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - y'_n)^2 \tag{3.12}$$

$$\epsilon_{MAE} = max(|\frac{y_n - y'_n}{y_n}|) \tag{3.13}$$

In the formula, N represents the number of predicted points; y_n is the true value of the nth predicted point; y'_n is the predicted value of the nth prediction point [18,19].

3.6. RF-DL-RST prediction model. The RF-DL-RST model framework is shown in Figure 3.2. The author's goal is to make short-term predictions of electrical loads, and the input feature quantities include various factors such as weather and time, which differ from the predicted results (that is load data) in terms of dimensions, units, etc. Therefore, preprocessing of the predicted data is necessary[20].

There are many factors that affect the electricity load in a region, including weather, time, and policies. However, the prediction accuracy of DNN is not positively correlated with the input items. When there are too many input items, it not only causes the network structure to be complex, but also may degrade the model accuracy.

The author establishes a feature set for load forecasting. However, the author believes that the week date, workdays, and holidays in its time factor constitute duplicates, so the characteristic quantity of the week date is excluded. At the same time, considering the impact of epidemic lockdowns on social electricity consumption in recent years, the author also studied whether the day was under lockdown as a characteristic quantity. In addition, the author also added weather factors such as average temperature, average wind speed, sunrise time, and sunset time as characteristic variables. The specific predicted feature quantities are shown in Table 3.1.

Table 3.1: Predicted feature quantities

Influence factor	Feature	Meaning
Time factor	Month	From January to December
	Day	Specific dates of each month
	Weekday	Normal work, value 1
	Festival and holiday	Saturdays, Sundays, and other holidays, with a value of 0
	Day Hour	00:00-24:00
Weather factors	Maximum temperature	The highest temperature of the day, ℃
	Minimum temperature	The lowest temperature of the day, ℃
	Average temperature	Average temperature of the day, ℃
	Average relative humidity	Daily average humidity,%
	Weather conditions	Such as sunny, cloudy, rainy, snowy, etc
	Air quality	Air quality index
	Average wind speed	Daily average wind speed, m/s
	Sunrise time	Specific time
	Sunset time	Specific time
Policy factors	Is it under lockdown or not	When affected by epidemics or natural disasters, take 1, otherwise take 0

3.7. Experimental Analysis. The author used load data from a certain regional power grid from October 28, 2022 to February 4, 2023 to simulate and verify the RF-DL-RST prediction model. In order to verify the superiority of the RF-DL-RST model, two comparative models are set up, among which: Comparative model 1 is the RF-DL model without RST correction part; Comparison Model 2 is a DL-RST model without RF feature selection. The selection of relevant parameters for the three models is consistent.

4. Results and Discussion.

4.1. Key feature extraction for load forecasting. Rank the importance of the predicted feature quantities selected in Table 3.1. The Random Forest (RF) model is configured with 500 decision trees and utilizes 3 split features. The dataset is divided into training and testing sets at a ratio of 9:1. Figure 4.1 illustrates the analysis findings regarding the importance of various features.

From Figure 4.1, it can be seen that after sorting the 15 feature quantities in Table 1 in order of importance scores from low to high, the 8 feature quantities of the day, including hours, minimum temperature, average temperature, weather conditions, holidays, workdays, sunrise time, and whether they are under control, have higher scores. Therefore, they are used as input items for the DNN model.

4.2. Deep learning training. Train the DNN model using the 8 key feature quantities and historical load data selected by RF as input and output items, respectively. The number of input layer nodes in DNN is 8, and the number of output layer nodes is 1. Set DNN to have 3 hidden layers with 40, 30, and 20 nodes respectively, and activate ReLU function; The ratio of training set to test set is 9:1, and the training frequency is 200 times. During the iteration process, the mean square error of the predicted values varies with the number of training iterations, as shown in Figure 4.2. It can be seen that the mean square error begins to converge at around 150 training iterations and continuously approaches the value of 975MW^2.

4.3. RST correction. According to equations 3.8-3.12, calculate the conditional attributes C={a,b,c}, as well as the decision attribute S before t, in order to obtain the rough set information system. Given the requirements of rough set theory for processing data, the encoding rule for the conditional attribute C={a,b,c} is set as follows:

$$C = \{a \in [1,6], b \in [1,3], c \in [1,6] | a, b, c \in Z\} \tag{4.1}$$

From this, the corrected load forecasting data can be calculated.

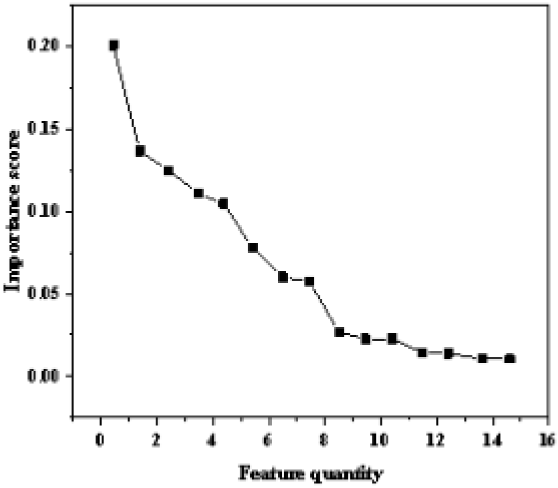


Fig. 4.1: Analysis results of feature importance in random forest algorithm

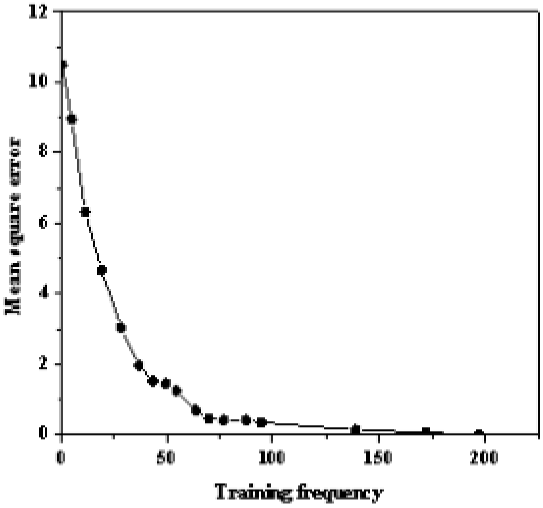


Fig. 4.2: Curve of Mean Square Error of Predictions as a Function of Training Times

Figure 4.3 shows the actual load on February 5, 2023 and the predicted load curve before and after RST correction. It can be seen that the predicted load curve after RST correction is basically between the actual load curve and the predicted load curve without RST correction, and is closer to the actual load curve.

4.4. Comparative analysis. According to equations 3.13 and 3.14, the evaluation indicators for the predicted results can be calculated. The comparison of indicators between RF-DL-RST model and RF-DL, DL-RST models is shown in Table 4.1.

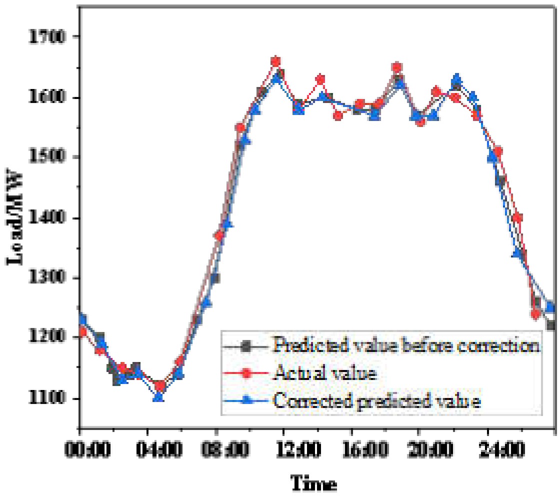


Fig. 4.3: Comparison of actual load and predicted load curve before and after RST correction

Table 4.1: Comparison of Indicators for Three Models

Model	DL training time/s	MSE MW^2	MAE/%
RF-DL-RST	96.28	680.32	4.02
RF-DL	96.28	974.64	5.76
DL-RST	107.20	865.83	4.72

From Table 4.1, it can be seen that compared with the RF-DL model, the MSE index of the RFDL-RST model has decreased by 30.187%, and the overall prediction results are closer to the true values. The MAE index has also decreased from 5.76% to 4.02%. During special periods of significant load changes such as 07:00-08:00 (rapid increase in load) and 22:00-23:00 (rapid decrease in load), the prediction accuracy has greatly improved. In addition, compared with the DL-RST model, the MAE and MSE indicators of the RF-DL-RST model were reduced by 15.210% and 21.414%, respectively, and the DL training time of the RF-DL-RST model was shortened by 10.175%, indicating that simplifying the DL input feature quantity through the RF model can improve the load forecasting effect. Based on the above analysis, it can be concluded that the RF-DL-RST model has significantly better prediction results, which verifies the effectiveness of the author's prediction model.

5. Conclusion. The author suggests a design for an electrical load forecasting system leveraging deep learning algorithms. Specifically, for short-term load forecasting, they introduce the RF-DL-RST model, which combines the random forest algorithm with rough set theory. Through example calculation and analysis, the following conclusions can be drawn: 1) By evaluating the importance of factors affecting load through RF, the model calculation time is shortened and the accuracy of prediction is improved. 2) By modifying the model results through RST and establishing evaluation models from both global and local perspectives, the effectiveness of the method was verified, greatly improving the accuracy of predicting load sudden changes.

6. Acknowledgement. Institute of Education Science, Chinese Academy of Management Science Research on the Cultivation of Students' Logical Reasoning from the Perspective of Core Literacy. ZGYJKS4090KT

REFERENCES

[1] Zhang, R., Yu, M., & Zhang, C. (2022). A similar day based short term load forecasting method using wavelet transform and lstm. IEEJ Transactions on Electrical and Electronic Engineering, 17(4), 506-513.

[2] Hao, L., Linghua, Z., Cheng, T., & Chenyang, Z. (2023). Short-term load forecasting model based on gated recurrent unit and multi-head attention, 30(3), 25-31.

[3] Wang, H., Zhang, N., Du, E., Yan, J., Han, S., & Liu, Y. (2022). A comprehensive review for wind,solar,and electrical load forecasting methods, 5(1), 22.

[4] Agyemang, F., Yamoah, S., & Debrah, S. K. (2022). Pseudocritical rapid energy dissipation analysis of base-load electrical demand reduction on nuclear steam supply system, 12(2), 19.

[5] Liu Bairu. (2022). Integration,coordination and empowerment-view the keywords of the"14th five-year plan"on photovoltaic development from the perspective of china's modern energy system planning, 29(2), 41-45.

[6] LI Xiao, & LU Xianling. (2022). Ethod for forecasting short-term power load based on dual-stage attention mechanism and gated recurrent unit network. Computer Engineering, 48(2), 291-296,305.

[7] Xu, L. H., Huang, C. Z., Wang, Z., Liu, H. L., Huang, S. Q., & Wang, J. (2024). Novel intelligent reasoning system for tool wear prediction and parameter optimization in intelligent milling, 12(1), 76-93.

[8] Sasidharan, M. P., Kinattingal, S., & Simon, S. P. (2023). Comparative analysis of deep learning models for electric vehicle charging load forecasting. Journal of The Institution of Engineers (India), Series B. Electrical eingineering, electronics and telecommunication engineering, computer engineering, 43(03), 747-761.

[9] Jin, X. B., Zheng, W. Z., Kong, J. L., Wang, X. Y., & Lin, S. (2021). Deep-learning forecasting method for electric power load via attention-based encoder-decoder with bayesian optimization. Energies, 14(6), 1596.

[10] Zhuo, C., Long-Xiang, S., & University, Z. (2018). Short-term electrical load forecasting based on deep learning lstm networks. Electronic Technology, 13(11), 249-267.

[11] Yan, X., & Zhu, H. (2023). A kernel-free fuzzy support vector machine with universum. Journal of Industrial and Management Optimization, 19(1), 282-299.

[12] Pengnian Qi,Yulun Liao,Biao Qin. (2023). Survey on deep learning for chinese named entity recognition. Journal of Chinese Computer Systems, 44(9), 1857-1868.

[13] Liu, P., Ahmad, S., Abdullah, S., & Mohammed M. Al-Shomrani. (2022). A new approach to three-way decisions making based on fractional fuzzy decision-theoretical rough set. International Journal of Intelligent Systems, 37(3), 2428-2457.

[14] Tie, J., Lei, X., & Pan, Y. (2022). Metabolite-disease association prediction algorithm combining deepwalk and random forest. Tsinghua Science and Technology, 27(1), 58-67.

[15] Yang, L. I., Wang, Q. Y., Tian, Q. H., Qi, A. N., Yang, Y. T., & Zhang, J. C., et al. (2024). Prediction of renal function by urinary lead and cadmium — based on classification decision tree and logistic regression model*. Biomedical and Environmental Sciences, 37(3), 331-335.

[16] Lyu, J., Bi, D. J., Liu, B., Yi, G., Zheng, X. P., & Li, X. F., et al. (2023). Compressive near-field millimeter wave imaging algorithm based on gini index and total variation mixed regularization, 21(1), 65-74.

[17] Tie, J., Lei, X., & Pan, Y. (2022). Metabolite-disease association prediction algorithm combining deepwalk and random forest. Tsinghua Science and Technology, 27(1), 58-67.

[18] Du, C., Du, C., Huang, L., Wang, H., & He, H. (2022). Structured neural decoding with multitask transfer learning of deep neural network representations. IEEE transactions on neural networks and learning systems, 33(2), 600-614.

[19] Kong, Q., & Chang, X. (2022). Rough set model based on variable universe, 7(3), 9.

[20] Al-Tameemi, I. K. S., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2023). Multi-model fusion framework using deep learning for visual-textual sentiment classification, 76(8), 2145-2177.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: May 11, 2024
Accepted: Jun 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, `http://www.scpe.org`
© 2025 SCPE. Volume 26, Issues 3, pp. 1386–1394, DOI 10.12694/scpe.v26i3.3901

CONSTRUCTION OF INFORMATION MANAGEMENT MODEL FOR COLLEGE STUDENTS BASED ON DEEP LEARNING ALGORITHMS AND DATA COLLECTION

LIN ZHU*

Abstract. In order to improve the efficiency of college student management and provide effective tools and assistants for college student management staff, the author proposes the construction of an information technology model for college student management based on deep learning algorithms and data collection. Firstly, use the FasterR-CNN model to detect the heads of personnel in the laboratory, Then, based on the output results of model detection, use the IoU algorithm to filter out duplicate detected targets, Finally, a coordinate based positioning method is used to determine whether there are people on each workbench in the laboratory, and the corresponding data is stored in the database. The main functions of this system include: (1) Real-time video monitoring and remote management of the laboratory, (2) Timed automatic photography detection and data collection provide data support for quantitative management in the laboratory, (3) Query and visualization of data on changes in laboratory personnel. The experimental findings demonstrate that our proposed model excels with an F1 Score exceeding 91%, showcasing robust generalization across detection confidence levels ranging from 50% to 99%. Notably, at a detection confidence of 96%, our model achieves its peak performance with an impressive F1 Score of 95.7%. This underscores the model's exceptional detection capabilities. Leveraging Faster R-CNN and IoU optimization, our laboratory personnel statistics and management system offer real-time personnel tracking and remote management functionalities tailored for office environments.

Key words: Convolutional neural network, Object detection, College student management, Personnel statistics, Management informatization

1. Introduction. The swift advancement of information technology has posed unparalleled challenges to university student management, yet it has also presented vast opportunities for enhancing student management practices. Especially regarding the precision and pertinence of student management work, it is currently the biggest challenge faced by traditional student management personnel. The development of information technology has brought good news to solve this problem [1]. It is precisely with such questions that analysis and research can bring innovative results to the model of student management in universities. Finally, with the strong utilization of information technology advantages, we will promote the healthy and stable development of student management in universities [2]. It must be acknowledged that the advent of the information age has brought tremendous convenience to people's lives and work. To some extent, the integration of information technology has undeniably enhanced the efficiency of our daily lives and professional endeavors. For university staff involved in student management, leveraging information technology has provided firsthand experience of its convenience and rapidity. Traditionally, communication and interaction between teachers and students primarily occurred through face-to-face interactions in the realm of student management. With the application of information technology, real-time video communication has become a reality [3]. Whether it's issues related to student learning or accommodation, many problems can be solved as quickly as possible through information-based communication methods. Especially in case of unexpected situations, the application of information technology highlights its advantages in terms of timeliness [4].

As the representative of information technology, the popularity of the Internet and mobile Internet has also had an increasingly profound impact on contemporary college students. Especially in the development of thinking, the Internet and mobile Internet have brought many new ideas and mysteries to college students, making those engaged in college student management clearly feel how backward the efficiency of traditional student management is [5]. Therefore, personnel engaged in the management of college students can learn and improve their knowledge and skills in student management through the help of the Internet and mobile Internet, a huge resource pool. Therefore, the application of information technology will greatly improve the efficiency of

*Student Affairs Section, Zibo Vocational Institute, Zibo, Shandong, 255300, China (`lindazhuzhu1981@163.com`)

university student management and become a powerful tool and assistant for university student management staff [6].

2. Literature Review. There is relatively little specialized research on big data in student management in universities, and more research is focused on the overall management of universities, mainly in the following two aspects. On the one hand, starting from the strategic thinking of empowering university management with big data, we aim to enhance the understanding of big data in university management. Mok, K. et al. have introduced a novel approach to signal management tailored for the comprehensive processing demands of smart cities. This method integrates deep learning and simulation techniques to address the challenge of handling large-scale data volumes effectively. Initially, the system undergoes offline training using deep learning networks based on computer vision, enabling the detection of various vehicle types. Training data, amassed from diverse city or country settings, facilitates a one-time training process for the deep networks. Subsequently, for each intersection requiring traffic flow prediction, a minimal dataset is gathered to construct a computer simulation model for localized traffic flow estimation. Finally, an adaptive traffic light management algorithm emerges through the fusion of deep learning-driven traffic monitoring systems and optimized simulation outcomes. This versatile approach offers seamless applicability across different intersections, requiring only minimal traffic data collection for each new intersection [7]. Yang, H. et al. have explored the integration of data mining technology within college student education management information systems, offering insights and strategies for enhanced student information management. Their work delves into research methodologies for employing data mining techniques within these systems, particularly focusing on educational data. They detail the implementation of the K-means and fuzzy C-means clustering algorithms, culminating in the development of a robust data mining system. Performance evaluation reveals an average algorithm processing time of 1.92 seconds within the system, ensuring a seamless user experience for education administrators [8]. Fan, J. et al. have primarily focused on leveraging data mining technology to advance university information management systems (IMS). Their investigation delves into the process of data mining (DM) and its application, particularly in association rule mining, within the realm of university IMS development. Their findings reveal a comprehensive course coverage, with specific parameters set for transaction support count and confidence level. Notably, the research highlights a tendency among students to select multiple courses in conjunction with one another. The integration of DM theory within university informatization initiatives is poised to significantly enhance the data analysis capabilities of management personnel, consequently elevating their proficiency in administration [9].

For the common office scenario of fixed indoor personnel and fixed workstations, the author takes ordinary university laboratories as an example and proposes an indoor personnel statistics method based on Faster R-CNN and Intersection over Union (IoU) optimization. This method utilizes deep neural networks to extract head features from images, resulting in higher detection accuracy; In addition, a coordinate based positioning method has been proposed, which can accurately determine whether there are people on each workbench in the laboratory. Finally, a laboratory personnel statistics and management system was designed and developed using the trained detection model, which effectively achieved remote, automated, and intelligent management of the laboratory. The experimental results indicate that the system can be applied to common indoor office scenarios.

3. Research Methods. Fast R-CNN is a rapid target detection method based on regional convolutional network (Region-based Convolutional Network). As an improvement of the R-CNN model, Fast R-CNN improves detection speed, but like R-CNN, it uses Selective Search (SS) method to extract candidate target regions (Proposals) from images. Therefore, there are still problems such as cumbersome detection steps, high time and memory consumption. FasterR-CNN introduces a Region Proposal Network (RPN) in the model to extract candidate target regions, achieving feature sharing in convolutional layers and greatly improving the generation speed of candidate target regions [10]. The Faster R-CNN network structure mainly consists of RPN and FastR-CNN detectors, where the input of RPN is image features extracted through a series of convolutions.

3.1. Feature extraction network. To illustrate, conventional deep neural networks like AlexNet, VGGNet, and GoogLeNet have the capacity to enhance image feature extraction by augmenting the number of network layers. However, when deep networks reach a point of convergence, elevating the number of layers may lead to a phenomenon termed ”degradation,” wherein the network’s detection accuracy plateaus or diminishes.

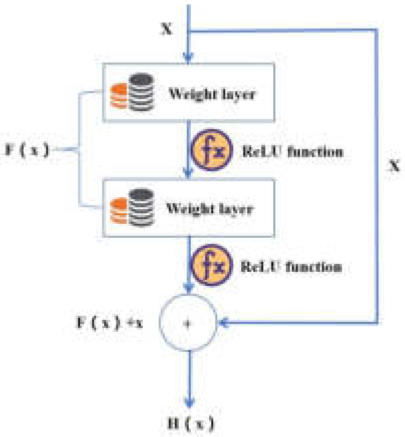


Fig. 3.1: Unit structure of residual network

Residual Neural Network (ResNet) can effectively solve the phenomenon of network degradation and has better image feature learning ability. Therefore, the author chose the residual network as the feature extraction network for Faster R-CNN [11].

The unit structure of the residual network is shown in Figure 3.1. Assuming that the original mapping output of the network units is $H(x)$, that is $H(x) = F(x) + x$. then $F(x) = H(x) - x$. Therefore, each convolutional output of the deep network will become a fitting residual. It can be simply understood that residual networks add some ”cross layer connections” in traditional deep convolutional networks (x in Figure 3.1), when the training error increases with the depth of the network, the residual network will skip certain convolutional layers and directly input the original data into the subsequent convolutional layers, which not only ensures the integrity of data transmission but also relatively reduces the training error and reduces the difficulty of deep network training [12].

3.2. Regional recommendation network. Traditional candidate target region extraction methods suffer from time-consuming issues, such as the sliding window and image pyramid used in Adaboost, and the SS used in R-CNN and Fast R-CNN. The RPN used by Faster R-CNN embeds the extraction of candidate target regions into the network and improves the generation speed of candidate target regions by sharing convolutional layer feature parameters. The author combines the actual pixel size of the target area and in order to obtain multi-scale detection boxes, RPN uses a 3x3 convolutional kernel to slide on the feature map output by the feature extraction network, and maps the region corresponding to the center of the convolutional kernel back to the original input image, generating a total of 12 anchors with 4 scales $\{16^{0.5}, 16, 16^{1.5}, 16^2\}$ and 3 aspect ratios {05,1,2}. Therefore, there are 12 suggested regions corresponding to the center of the convolution kernel in each sliding window. RPN is a fully convolutional network that inputs the original image convolutional feature map output by the feature extraction network. The suggested region corresponding to each anchor point is convolved through an intermediate layer to output a 512 dimensional feature vector, which is then fed into the classification layer and position regression layer, respectively. Among them, the classification layer outputs the classification information of the target in the corresponding anchor point, including the confidence level of the background and the confidence level of the target category; The position regression layer outputs the position information of the target in the anchor point, including the center point coordinates, length, and height of the target area. Finally, using the Non Maximum Suppression (NMS) algorithm, the candidate target regions are filtered based on the classification and position information of all anchor points, resulting in 2000 high-quality target candidate regions [13,14].

3.3. Fast R-CNN detection network. Once the Region Proposal Network (RPN) generates candidate regions of interest, they undergo further refinement through the Fast R-CNN detector for accurate classification and coordinate regression. To address the diverse sizes of these candidate regions, they are directed to the Region

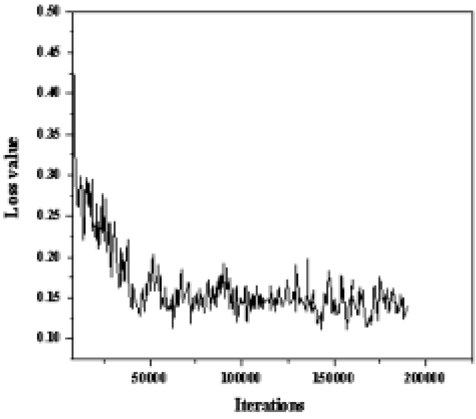


Fig. 4.1: Loss values during training process

of Interest (RoI) pooling layer, where they are resized uniformly for streamlined processing. The RoI pooling layer combines feature maps and target candidate regions for coordinate mapping, outputting a fixed size target candidate region. Subsequently, these target candidate regions are sent to the Fast R-CNN detector for training, obtaining the final detection results including classification information and coordinate information [15].

4. Result analysis.

4.1. Data generation and training. The author's experimental data was collected using a monocular camera located at the top of the laboratory. The image captured by the top camera shows a significant amount of occlusion in various parts of the human body. Therefore, the human head is selected as the detection target to determine the number and distribution of personnel in the experiment. A total of about 6000 original images were collected, and after flipping and symmetry, the dataset was expanded to about 24000 images. The image size is uniformly 1 510 x 860 pixels, and the number of people in each image ranges from 1 to 10. Randomly divide the dataset into training and testing sets in a ratio of 10:1 [16]. The author's experimental environment is Windows 10, GeForce GTX 1080Ti, and the network model is implemented using the mainstream deep learning framework TensorFlow. In the overall model architecture, ResNet101 serves as the feature extraction backbone. The training process employs a batch size of 4, initializing the learning rate at 0.0003. After 40,000 iterations, the learning rate decreases to 0.00003, followed by a further reduction to 0.000003 after 80,000 iterations, concluding with a total of 200,000 iterations [17].

4.2. Result Analysis. Following the training phase, the Faster R-CNN model attained an impressive mean average precision of 98.49% when evaluated on the test dataset. The training progression, as depicted in Figure 4.1, illustrates the loss curve over the course of training.

Figure 4.1 illustrates that the model's loss value stabilized around 0.15 after 180,000 iterations, indicating convergence. The exceptional detection performance of the model can be attributed to:

1. The scene background in the laboratory is single, with low personnel mobility, less personnel and background changes, and more prominent image features;
2. There are many training data samples, and the training set contains more than 20000 images, totaling about 70000 labeled human head samples;
3. For targets of different scales, a total of 12 anchor points with 4 scales and 3 aspect ratios are used, which can effectively detect targets of different scales;
4. The model utilizes RPN to generate high-quality target candidate regions, providing high-quality training data for subsequent Fast R-CNN networks.

Table 4.1: Four scenarios for model detection

Situation	Detected as the target	Detected that it is not the target
Actually, it's the goal	TP (really)	FN (False No)
Actually, it's not the target	FP (False)	TN (True or False)

In order to further study the generalization ability of the model, that is, its detection performance in actual scenes, 105 images were collected from the images captured by the camera as an incremental test set to test the detection performance of the model under different confidence levels. The most commonly used evaluation indicators for detection models are accuracy and recall. There are four situations when calling the model for detection: 1) Actually, it is the target, and detection considers it as the target; 2) In fact, it is the target, and detection considers it not the target; 3) In fact, it is not a target, and detection considers it as a target; 4) Actually, it is not the target, and the detection assumes it is not the target [18].

The four possible scenarios for the model to detect targets are shown in Table 4.1.

Therefore, the definitions of precision and recall can be given as follows: 4.1,4.2:

$$P = TP/(TP + FP) \tag{4.1}$$

$$R = TP/(TP + FN) \tag{4.2}$$

Among them, P is the model accuracy, and R is the model recall. Accuracy refers to how much of the detection results provided by the model are correct, while recall refers to how many actually correct targets have been detected. These two indicators usually have a trade-off. In order to comprehensively consider these two indicators, a new evaluation indicator is introduced, which is the weighted harmonic mean F-Score of accuracy and recall. The following equation 4.3:

$$F - Score = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \tag{4.3}$$

Among them, β for harmonic parameters, when $\beta > 1$, accuracy is considered more important, the author believes that recall and accuracy are equally important, that is, taking β=1. Therefore, the author's weighted harmonic mean is F1 Score, as follows:

$$F1 - Score = \frac{2 * P * R}{P + R} \tag{4.4}$$

The incremental test set consists of 105 images, with a total of 445 actual targets. Test the model using an incremental test set and calculate the F1 Score at different detection confidence levels, as shown in Figure 4.2.

From Figure 4.2, it can be seen that the model has a high F1 Score with a detection confidence level of 50% to 99%, and the F1 Score is greater than 91%, indicating that the model has strong generalization ability. At the same time, when the detection confidence is 96%, F1 Score reaches the highest level, reaching 95.7%, indicating that the model has the best detection performance at this time, that is, when the detection confidence is 96%, the model's generalization ability is the best.

During the incremental testing process, the detection time was 28.89 seconds, with an average detection speed of 275.1 ms per image, which is much lower than the detection speed in the middle, indicating that although the model implemented by the author has high detection accuracy, the detection speed is difficult to achieve the goal of real-time detection of video streams. In the field of object detection, there are usually three methods that can improve detection speed:

1. Reduce the size of the input image. This method is suitable for scenes with small monitoring areas. The author's input image size is 1 510 x 860 pixels, which only covers the experimental monitoring area. Therefore, this method is not applicable.

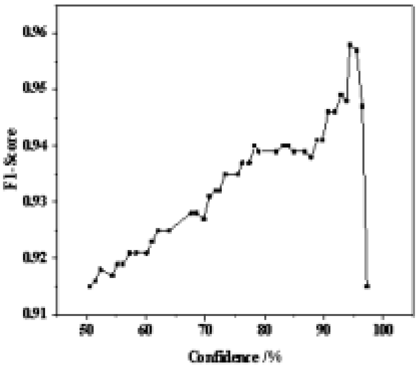


Fig. 4.2: Under different confidence levels

2. Using hardware devices with more powerful computing performance. This method requires a significant amount of capital and is therefore more suitable for the industrial sector.
3. 3) Utilize a more streamlined detection model based on network structure. This method usually sacrifices detection accuracy to a certain extent, such as SSD, making it more suitable for application areas that do not require high detection accuracy and high detection speed.

In the laboratory, personnel are mostly sitting and working, with low mobility. It is unnecessary to blindly pursue detection speed in this scenario, so the author did not use the three methods mentioned above. Instead, the system was designed to automatically detect every minute. Therefore, the detection speed of the model meets the design requirements of the system. For common office scenarios, a design that checks every minute is sufficient to provide reliable data for scientific management by managers [19].

4.3. IoU optimization. The accuracy and generalization ability of the model implemented by the author are quite outstanding, and the detection results are relatively good. The detection model detects a human head into two. In response to this situation, the author utilized the IoU algorithm for further optimization. IoU refers to the overlap rate of two detection boxes with overlapping areas, which is the ratio of the intersection and union between these two detection boxes. As shown in Figure 4.3, there is an overlapping area between Box A and Box B. Among them, S (A), S (B), and S (C) represent the areas of boxes A, B, and C, respectively.

After using the Faster R-CNN object detection model to detect the target image, all detection boxes containing position information output from the detection are input into the IoU algorithm. Among them, number is the number of targets detected by the model, which is the number of personnel in the laboratory output by the model detection. The final output N is the actual number of people in the laboratory filtered by IoU.

4.4. Personnel positioning. In response to the characteristics of low personnel mobility, single environment, and relatively fixed personnel positions in the laboratory, the author proposes a coordinate positioning method to determine whether there are people on each workbench and store the corresponding data in a database, providing reliable data support for scientific management of the laboratory. Divide the monitoring area into 12 rectangular areas in advance, representing the workstations within each area. Firstly, the center of mass of the human head is determined using the position information of the personnel target detected by the model, and then discrimination is carried out one by one. If the center of mass falls in which area, it is considered that there is a person on the workbench in that area. It can be considered that there are people on the workbench in areas 2, 3, 4, 6, and 8 [20].

4.5. System Implementation and Display. The system is developed using the open-source web development framework Django, and has two main functional modules: the system's historical data query and

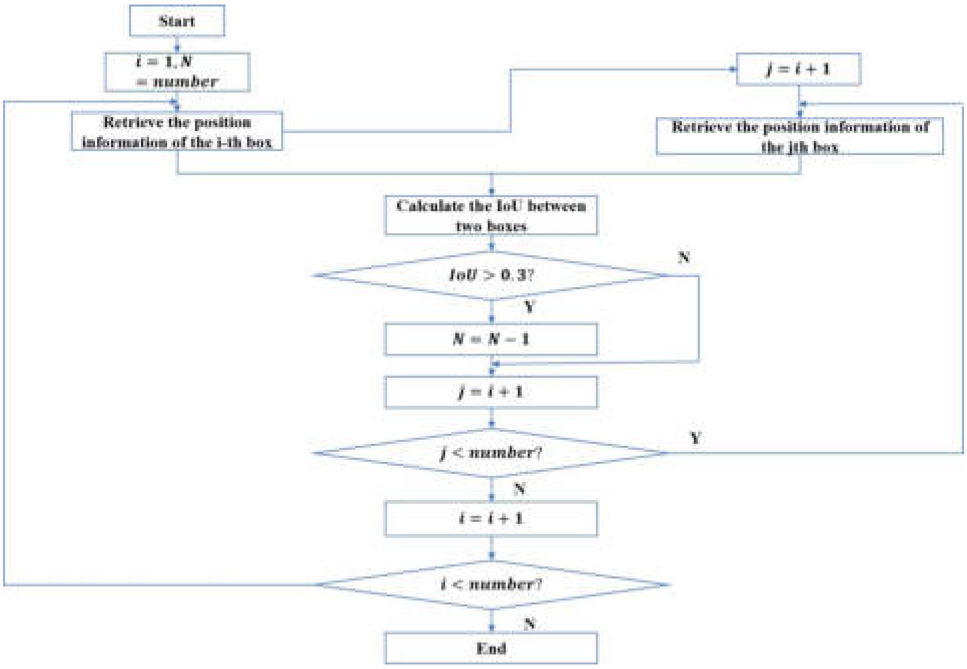
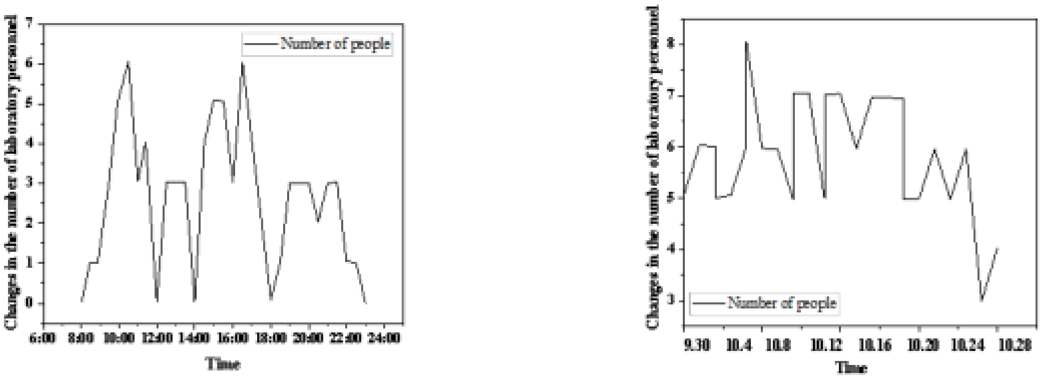

Start
$i = 1, N = number$
Retrieve the position information of the i-th box
$j = i + 1$
Retrieve the position information of the jth box
Calculate the IoU between two boxes
$IoU > 0.3$?
N
Y
$N = N - 1$
$j = i + 1$
$j < number$?
Y
N
$i = i + 1$
$i < number$?
N
End

Fig. 4.3: IoU Algorithm Process



(a) Changes in daily population

(b) Changes in the number of people over a period of time

Fig. 4.4: Changes in the number of laboratory personnel

display module, and the real-time video monitoring and detection module. The system is developed based on B/S mode and has remote management function. Authorized users can access the system within the campus network by logging in through a PC browser. From 6:30 to 23:30 every day, the system server automatically captures a laboratory monitoring image every minute and calls the detection model to detect it; Then locate the personnel based on the test results to determine if each workbench is manned; Finally, store the corresponding data in the database for laboratory administrators to query.

Figure 4.4 shows the query page for daily changes in the number of people in the laboratory and changes in the number of people over a period of time. This system has been running stably in the laboratory for six months, and its promotion and application value has been verified [21].

5. Conclusion. In response to the characteristics of fixed personnel and fixed workstations in common office scenarios, taking ordinary university laboratories as an example, the author proposes an indoor personnel statistics method based on Faster R-CNN and IoU optimization. The experimental results show that the proposed method has good detection accuracy. Then, based on the detection results, a coordinate positioning method is used to determine whether each indoor workbench is occupied. Finally, a laboratory personnel statistics and management system was developed using the Django framework, achieving remote, automatic, and intelligent management of the laboratory. However, the system developed by the author has the problem of personnel positioning relying on the premise that personnel positions are relatively fixed. When personnel positions move, the system cannot make accurate judgments. Therefore, further research on target tracking algorithms between video frames will be carried out, while ensuring detection accuracy. By drawing the movement trajectory of personnel, dynamic positioning of personnel will be achieved.

6. Acknowledgement. The research is supported by: This work was supported by the Natural Science Foundation of Province, China (Grants No.2019JJ40097), by the Research Foundation of Hunan University of Science and Engineering, China (Grant No.17XKY068) and by the construct program of applied characteristic discipline in Hunan University of Science and Engineering.

REFERENCES

[1] Qu, J. (2021). Research on mobile learning in a teaching information service system based on a big data driven environment. Education and Information Technologies, 26(5), 6183-6201.

[2] Li, X., Liu, H., Wang, W., Zheng, Y., Lv, H., & Lv, Z. (2022). Big data analysis of the internet of things in the digital twins of smart city based on deep learning. Future Generation Computer Systems, 128, 167-177.

[3] Bi, H., Liu, J., & Kato, N. (2021). Deep learning-based privacy preservation and data analytics for IoT enabled healthcare. IEEE Transactions on Industrial Informatics, 18(7), 4798-4807.

[4] Kor, M., Yitmen, I., & Alizadehsalehi, S. (2023). An investigation for integration of deep learning and digital twins towards Construction 4.0. Smart and Sustainable Built Environment, 12(3), 461-487.

[5] Jasti, V. D. P., Zamani, A. S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., ... & Kaliyaperumal, K. (2022). Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis. Security and communication networks, 2022, 1-7.

[6] Hu, J. (2021). Teaching evaluation system by use of machine learning and artificial intelligence methods. International Journal of Emerging Technologies in Learning (iJET), 16(5), 87-101.

[7] Mok, K. , & Zhang, L. . (2024). Adaptive traffic signal management method combining deep learning and simulation. Multimedia tools and applications(5), 83, 639-653.

[8] Yang, H. , & Zhang, W. . (2022). Data mining in college student education management information system. Int. J. Embed. Syst., 15, 279-287.

[9] Fan, J. , Zhang, M. , Sharma, A. , & Kukkar, A. . (2022). Data mining applications in university information management system development. Journal of Intelligent Systems, 31(1), 207-220.

[10] Dhillon, A., & Verma, G. K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 9(2), 85-112.

[11] Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. Journal of cognitive neuroscience, 33(10), 2017-2031.

[12] Tian, Y. (2020). Artificial intelligence image recognition method based on convolutional neural network algorithm. Ieee Access, 8, 125731-125744.

[13] Zhou, D. X. (2020). Theory of deep convolutional neural networks: Downsampling. Neural Networks, 124, 319-327.

[14] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks?. Advances in neural information processing systems, 34, 12116-12128.

[15] Ilesanmi, A. E., & Ilesanmi, T. O. (2021). Methods for image denoising using convolutional neural network: a review. Complex & Intelligent Systems, 7(5), 2179-2198.

[16] Wu, X., Sahoo, D., & Hoi, S. C. (2020). Recent advances in deep learning for object detection. Neurocomputing, 396, 39-64.

[17] Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.

[18] Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2020). Imbalance problems in object detection: A review. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3388-3415.

[19] Dhillon, A., & Verma, G. K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 9(2), 85-112.

[20] Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: Challenges, architectural successors, datasets and applications. multimedia Tools and Applications, 82(6), 9243-9275.

[21] Pal, S. K., Pramanik, A., Maiti, J., & Mitra, P. (2021). Deep learning in multi-object detection and tracking: state of the art. Applied Intelligence, 51, 6400-6429.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: May 11, 2024
Accepted: Jun 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1395–1402, DOI 10.12694/scpe.v26i3.4102

PERSONALIZED OPTIMIZATION OF SPORTS TRAINING PLANS BASED ON BIG DATA AND INTELLIGENT COMPUTING

ZHONG DING*

Abstract. In order to effectively improve the athletic performance of athletes and make their training more systematic, scientific, and standardized, the author proposes a personalized optimization study of sports training plans based on big data and intelligent computing. Based on the research on big data and intelligent computing, the author designs a human-computer interaction system using association rule mining algorithms, data fusion processing in sports training decision support systems, and improved Apriori algorithm frequent association rules. A sports training plan leveraging big data and intelligent computing was developed through experiments. The findings reveal that the author's enhanced Apriori algorithm exhibits a reduced reaction time when subjected to a minimum support threshold. Specifically, with a minimum support of 3.5, the execution time is under one second. This demonstrates that the improved Apriori algorithm can efficiently facilitate decision support for sports training regimes, offering valuable insights for athletes' physical conditioning. The research on personalized optimization of sports training programs, utilizing big data and intelligent computing, enables athletes to access real-time sports data, thereby enhancing their performance.

Key words: Big data, Intelligent computing, Sports, Training plan

1. Introduction. With the rapid progression of modern technology, the convergence of various disciplines has steadily increased. To effectively improve athletes' performance, incorporating computer technology support systems can make their training more systematic, scientific, and standardized. Presently, data mining technology is a major area of development in China and has been widely applied across many industries. By analyzing sports training methods using data mining techniques, we can develop decision-making systems that better optimize traditional sports training methods. The importance of digital physical training monitoring lies in its ability to monitor and analyze the training process in real-time and in detail. By configuring digital equipment, data such as exercise posture, muscle strength, and training energy range can be read. These data can not only help coaches and athletes provide timely feedback on speed, strength, endurance, agility, and coordination control abilities, but also enable the development of more scientific and personalized training plans based on these data [1]. Through digital technology, the training process of athletes can be monitored in real-time, problems can be identified in a timely manner, training plans can be adjusted, and training efficiency can be improved. Digital technology can monitor the training process of athletes in real-time, detect their physical condition in a timely manner, and prevent athletes from being injured due to overtraining [2].

Sports are an indispensable part of people's lives, and as future sports talents, the quality of their training is directly related to the development of the sports industry. However, traditional sports training methods have some problems, such as difficulty in evaluating training effectiveness and frequent sports injuries. Digital training monitoring utilizes modern technological methods to measure and analyze data in real-time during athletes' physical training, in order to monitor training quality and adjust the physical training process based on the data. For example, digital monitoring of strength training and resistance training, as well as other important aspects of physical fitness training, can be achieved through the intervention of big data technology to achieve digital and real-time monitoring. In high-level competitive training, ”no monitoring, no training” has become a new concept and criterion. The training team will conduct data-driven monitoring, feedback, and optimization of techniques, tactics, physical fitness, condition, field, and command throughout the entire process, in order to achieve the transformation from traditional experiential training to information-based, digital, and scientific training [3]. In addition, by monitoring the training of top athletes in various sports

*Linyi University, Shandong, 276000, China. (Corresponding author, dingzhongsh@163.com)

around the world and monitoring the daily physical fitness of the general public, a large amount of data has been collected for analysis and comparison, and a reliable data analysis framework and system have been designed. This approach not only enhances the precision and efficiency of physical training but also aids coaches in understanding the training effects, revising training plans, and scientifically managing the training process. Sports encompass a broad range of disciplines, including humanities, sports science, and sports social science. The rapid advancements in computers and information technology, particularly in artificial intelligence theory and data mining technology, have established a strong theoretical foundation for scientific training and the implementation of advanced training methods. As training data accumulates, traditional statistical analysis techniques may fall short in effectively analyzing the data, making it challenging to identify suitable patterns to describe the correlations. Data mining, however, offers optimization methods for uncovering scientific patterns and correlations within extensive and complex training data [4].

2. Literature Review. Through digital technology, the training effectiveness of athletes can be accurately evaluated, scientific training suggestions can be provided, and training quality can be improved. Researchers, both domestically and internationally, have carried out studies on decision support systems for sports training methods. Guan, L. K. et al. examined the development process of an intelligent decision support system for college sports using big data analysis. They integrated big data and artificial intelligence technologies into the creation of university sports decision support systems, proposing a system framework and its structural components. Finally, they analyzed the relevant key technologies, offering a reference for building similar systems [5]. Huang, Y. et al. developed an intelligent sports prediction analysis system by integrating the predictive capabilities of the particle swarm optimization algorithm in edge computing with traditional sports event prediction methods. Experimental results demonstrated that this intelligent sports event prediction and analysis system achieves higher accuracy in forecasting sports events compared to conventional prediction methods and better meets the interests of sports event enthusiasts [6]. Ai, X. B. et al. introduced an algorithm for intelligently integrating traditional ethnic sports and cultural resources through big data. They initially established a comprehensive dataset by defining the time decay period of weighted samples. Mining parameters were subsequently based on accurate values to enable in-depth exploration of the insights within traditional ethnic sports and cultural resources [7].

In order to address the above issues and provide a basis for scientific sports training, the author proposes a personalized optimization study of sports training programs based on big data and intelligent computing. Intelligent computing technology is a discipline that describes problem objects through specific mathematical models, making them operable, programmable, computable, and visual. It utilizes its parallelism, adaptability, and self-learning to mine patterns and discover knowledge from massive data in disciplines such as information, neurology, biology, and chemistry. Based on the research of previous scholars, this paper explores the application of big data and intelligent computing in sports training mode decision support systems.

In order to reduce the negative impact of current laser image pattern recognition methods on image applications, a laser image pattern recognition method based on big data analysis is proposed. Firstly, briefly explain the principles of big data analysis applied in the research process, Implement image analysis and training processing on the original laser image big data, Extract the features of laser images and perform pre segmentation on them to achieve pre-processing of the original laser images, Finally, the pattern recognition of laser images is completed using the regularized least squares method. Set up a simulation experiment to verify the application effect of this method. Using 10 types of laser images and a total of 1000 images as samples, set the preset target image recognition rate, recognition error rate, and image recognition consumption time as evaluation indicators for the simulation experiment. Apply this method for image pattern recognition analysis.

3. Research Methods.

3.1. Big Data and Intelligent Computing.

3.1.1. Application of Big Data and Intelligent Computing. The advent of the intelligent era is propelled by the collaborative advancements in big data and deep learning technologies. Artificial intelligence, conceptualized over six decades ago in 1956, continues to evolve, necessitating a strong integration with the real economy to leverage existing industrial frameworks and foundations. And every enterprise or region should be pragmatic, tailored to local conditions, and have a persistent attitude, so that artificial intelligence can truly

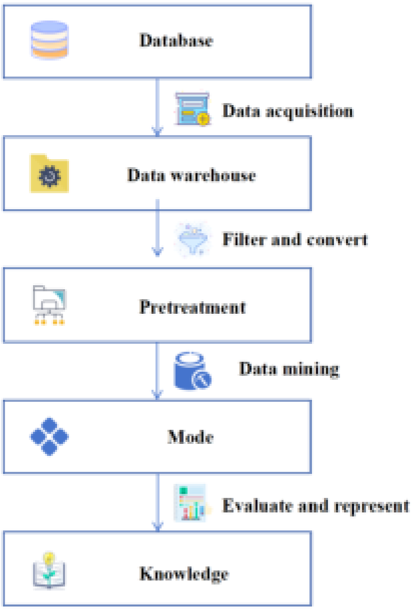


Fig. 3.1: Data Mining Process

lead social change. Artificial intelligence has played a lot of roles in various industries, and countries are also doing it. The future is an era of intelligence, where artificial intelligence technology influences various industries and fields. These fields are not independent, and they can cross integrate with each other. The upstream and downstream of the entire industry chain also influence each other. From the analysis of small data to data mining, to massive data mining, and now to big data mining, we have been exploring how to utilize the useful information hidden in data to solve problems [8]. Of course, in the process of data mining, it is necessary to deal with issues such as the distribution, volume, dimension, uncertainty of data, and differences in human cognition for the same data. The ultimate goal of using computer systems to process data is to complete human self-awareness, use computer systems to calculate, solve cognitive problems, and then solve problems. Human cognition has a habit of prioritizing large areas, seeing the big picture first and then the details in the middle. This is universal. Computer data processing ranges from fine to coarse, and multi granularity big data intelligent computing can be done using a mathematical model, which is what we call granularity problems. We classify items, and different classifications reflect different granularities; On the cloud model, connect human cognitive behavior with computer data processing behavior, create a collaborative computing model between humans and machines, and create a granular cognitive computing model. In various industries, we naturally apply this model concept to turn enterprise management into a multi granularity model and make intelligent decisions at different granularities.

3.1.2. Overview of Association Rule Mining Algorithms. Data mining technology, particularly the association rule algorithm like the Apriori algorithm, plays a pivotal role in analyzing complex datasets. Association rules are commonly classified based on the dimensions of data types they encompass. This classification typically divides rules into single-dimensional and multi-dimensional categories. In sports training for athletes, where numerous influencing factors are at play, the data collected often exceeds three dimensions. Therefore, when developing a decision support system for designing sports training modes using data mining techniques, it becomes crucial to consider multi-dimensional association rules. These rules are more intricate than single-dimensional ones, involving processes such as data preprocessing, mining, and pattern evaluation, as illustrated in Figure 3.1.

The preprocessing stage in the data mining process mainly involves collecting, processing, and transforming data, which takes the longest time throughout the entire data mining process; The data mining stage mainly analyzes the data in the preprocessing stage through selected association rules, neural network techniques, etc; In the pattern evaluation stage, the focus lies on presenting the insights derived from data mining to users. This can involve creating visual programs that allow for real-time viewing and analysis of the information obtained. When evaluating sports training modes with big data mining technology, it's crucial to seamlessly integrate relevant data into the sports evaluation decision support system. The author mainly uses neural network models to classify the extracted data features [9,11].

3.2. Data Fusion Processing of Sports Training Decision Support System.

3.2.1. Fusion clustering of decision information. The data feature identification function of the sports evaluation decision support system is defined by formula 3.1:

$$P_c = \sum_{i=1}^{n} \sum_{j=0}^{n} \alpha(i,j) P(i,j) \tag{3.1}$$

Formula 3.2 expresses the model relationship between sports training evaluation decision data and the distribution of cluster centers.

$$p_r = \frac{P_t}{(4\pi)^2 (\frac{d}{\lambda})^\gamma} [1 + \alpha^2 + 2\epsilon \cos(\frac{4\pi h^2}{d\lambda})] \tag{3.2}$$

Formula 3.3 represents the attribute categories of association criteria within the sports decision support system, guiding the feature recognition process based on various types of acquired data:

$$R_\beta X = U\{E \in U/R | c(E,X) \leqslant \beta\} \tag{3.3}$$

$$R_\beta X = U\{E \in U/R | c(E,X) \leqslant 1 - \beta\} \tag{3.4}$$

By mining the data of the sports training mode decision support system and extracting its association rules, the establishment of the system database model is achieved, and corresponding system design is carried out in combination with software development [12].

3.2.2. Design of human-computer interaction system. Improve the Apriori algorithm for frequent association rules in the design of human-computer interaction systems [13]. Its primary functions include: 1) Facilitating a user-friendly computer interaction system for decision-makers in sports training. This system enables checking each athlete's physical fitness indicators through front-end display settings and utilizes computers for efficient tracking and processing. 2) Visually display the operational status of the decision support system, allowing users to fully understand the data changes during the system's operation and make timely adjustments [14]. 3) Based on the output results of the system, targeted adjustments can be made to the training plan, and the simulation can be calculated to form the optimal training plan. 4) The human-computer interaction system also includes capabilities to rectify incorrect information and conduct initial verification and assessment of input data.

The training process mainly includes 5 stages, including student state diagnosis, training objectives, training plan, training plan, and goal completion evaluation, as shown in Figure 3.2. Among them, training analysis is a key link in sports training.

3.2.3. Decision Tree Based Data Mining Model. The rough set algorithm evaluates knowledge by approximating descriptions based on existing knowledge bases, aiming to eliminate redundant data during processing and achieve more precise decision outcomes. Traditional rough sets focus on evaluating and processing classified resource data, but additional data processing often necessitates discretization, which may result in information loss and data reduction [15,16]. Decision trees can classify data through a series of rules, and the classification rules represented by decision trees can be inferred from a set of irregular elements. In general, decision trees follow a recursive top-down approach where attribute values at internal nodes are compared to

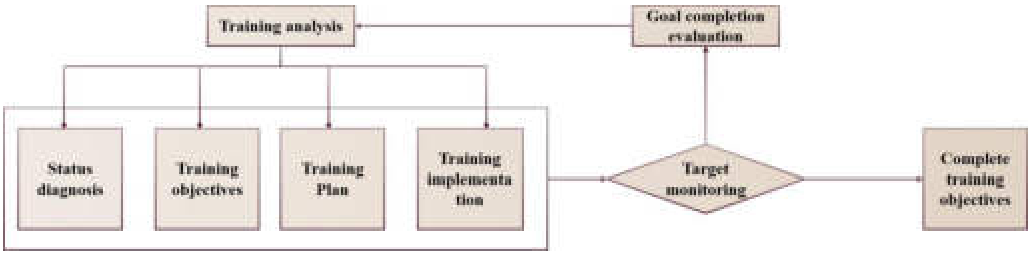


Fig. 3.2: Training Implementation Process

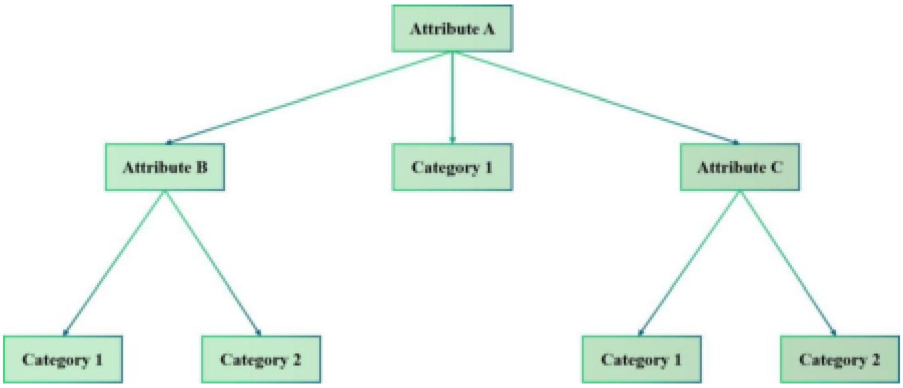


Fig. 3.3: Typical Decision Tree Composition

branch down based on different values, leading to leaf nodes that represent distinct classes. Thus, the path from the root to each leaf node forms a classification rule. Figure 3.3 illustrates a typical decision tree, comprising decision nodes, branch nodes, and leaf nodes. Each node corresponds to a non-categorical attribute, branches represent possible attribute values, and leaf nodes signify categories. Nodes in the middle of the tree are typically depicted as rectangles, while leaf nodes are shown as ellipses [17]. However, traditional decision trees are susceptible to issues like redundant branches due to noise and interference from abnormal data.

In order to address the above issues, Figure 3.4 shows the author's improved decision tree algorithm. Algorithms can be divided into two stages: learning and testing. In the learning phase, a top-down recursive approach is employed to train the parameters. Following this, the model and parameters are entered into the testing phase for validation and optimization [18]. This algorithm primarily involves two steps: first, generating the tree; and second, pruning the tree to eliminate data that may contain noise or anomalies.

3.3. Experimental research.

3.3.1. Development of data mining and training plans. The DSS of sports training model is based on the modern computer, and the computer and the programming language are used to simulate the performance of the athletes. The decision system provides a convenient way for players to access sports data in real time, and helps to guide and follow the training effectiveness. Based on the analysis of the DSS algorithms, the author integrates the DM technique into the system program, and a DSS for PE training is obtained. The main function modules of the system are the communication module, the program module, the database module and the data output module. Based on Conti-ki Bus, the data types are transferred and coordinated in DSS. In addition, the VIX integrated control technique is used to realize the overall control of the system. The DSS Data Perception component is based on a 6LoWPAN protocol stack. The WSN uses Atmel1284P as the main chip to process the whole Internet Network address assignment and management. Once the taskbar address is established, the system's human-computer interaction is facilitated through the TaskBasic interface program.

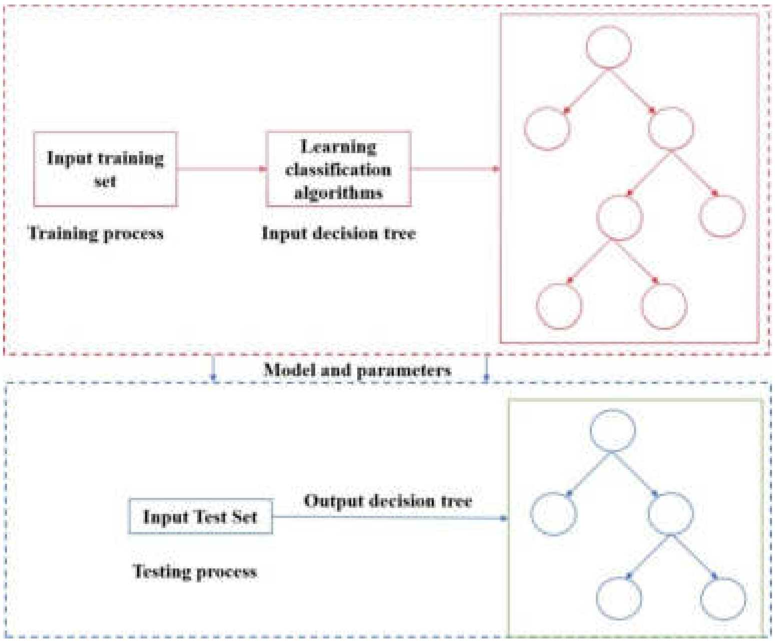


Fig. 3.4: Improved decision tree generation process

Table 4.1: Training related data of badminton teams in a certain area

types	Tid	Item	Avg
data	6511	226	47

Data quality is evaluated based on several factors, with the three most crucial being accuracy, completeness, and consistency. However, the experimental dataset contains errors, missing information, and inconsistencies, necessitating data preprocessing to enhance its quality and, consequently, the quality of data mining results. Additionally, various sports have distinct attributes for evaluation, such as time for track and field events, and attributes like score, hit rate, and duration for ball games. In order to effectively conduct data mining, the different values of each attribute can be mapped to a series of integers, and the values of attributes in that category can be replaced with integers.

4. Result analysis. In order to verify the validity of the modified Apriori algorithm, we compared the traditional Apriori algorithm, the DC Apriori algorithm and the modified Apriori algorithm. The experiments were primarily programmed using the Java language. The data set was composed of the training data of a district badminton team, as shown in Table 4.1.

In Table 4.1, Tid, Item, and Avg represent the specific training items, the total number of data items, and the average of the training sessions, respectively. Figures 4.1 and 4.2 illustrate the changes in the execution time of the system as the minimum support and the minimum confidence level increase.

Figure 4.1 demonstrates that the improved Apriori algorithm proposed by the author exhibits a shorter response time under minimum support conditions. Specifically, when the minimum support is set to 3.5, the execution time is less than one second, indicating superior algorithm performance. As shown in Figure 4.2, the improved Apriori algorithm performs better when the lowest confidence level is low. However, with the increase of the minimum confidence, the performance of the modified Apriori algorithm is reduced, and finally the same as that of the conventional Apriori algorithm. It is very important to improve the training efficiency of athletes and to optimize the support system of sports training. In this paper, the decision support of sports

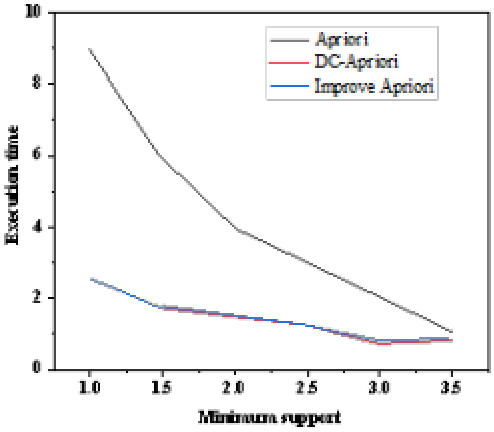


Fig. 4.1: Comparison of execution time under minimum support

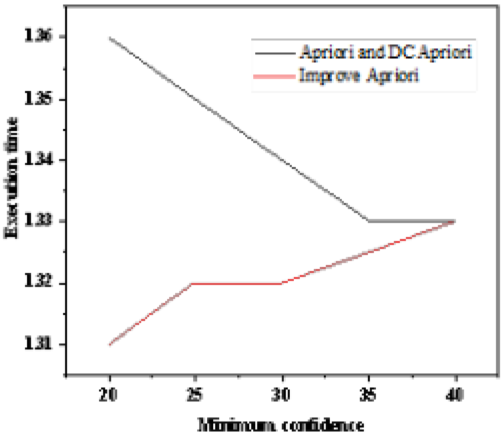


Fig. 4.2: Comparison of execution time under minimum confidence level

training model is studied by means of data mining, and an improved Apriori algorithm is put forward. This approach involves deep analysis and mining of system data to effectively extract information during athlete training, integrating sports evaluation decision data through relevant association rules. Moreover, the results of simulation and analysis on the traditional Apriori algorithm, DC Apriori algorithm and the modified Apriori algorithm prove that the improved Apriori algorithm can effectively support the decision of the sports training model.

5. Conclusion. The author addresses the need for scientific management and decision-making in the training process of sports athletes, combined with advanced big data and intelligent computing technology, and proposes an improved sports training mode decision support evaluation system. The author provides a detailed analysis of the characteristics of association rule algorithms and delves into their specific applications in data preprocessing, data mining, and pattern evaluation. Through these analyses, the system can achieve personalized optimization research on sports training programs, provide training suggestions based on scientific data for athletes, and thus improve training effectiveness and sports performance. The author not only enriches the scientific methods of sports training in theory, but also provides important references and support for athlete training in practice.

REFERENCES

[1] Sumer, Z. , & Lehn, R. C. V. . (2022). Data-centric development of lignin structure-solubility relationships in deep eutectic solvents using molecular simulations. ACS Sustainable Chemistry & Engineering(31), 10.

[2] Zhong, S. , & Hu, T. . (2022). A multi-attribute recognition method of vehicle's line-pressing in parking lot based on multi-task convolution neural network. International Journal of Information and Communication Technology, 20(3), 308-.

[3] (2023). Statement of retraction: sequential pattern data mining algorithm and blockchain technology for preparing a housing price prediction model. IETE Journal of Research, 69(11), 8515-8515.

[4] Deng, W. , & Zheng, H. . (2024). Construction of "online-offline" teaching and process evaluation system for digestive endoscopy by standardized training doctors. Open Journal of Social Sciences, 12(2), 10.

[5] Guan, L. K. . (2023). Research and Construction of University Sports Decision Support System Based on Extraction Algorithm and Big Data Analysis Technology. EAI International Conference, BigIoT-EDU. Springer, Cham.

[6] Huang, Y. , & Bai, Y. . (2023). Intelligent sports prediction analysis system based on edge computing of particle swarm optimization algorithm. IEEE consumer electronics magazine, 29(3), 496-502.

[7] Ai, X. B. . (2022). Intelligent integration algorithm of national traditional sports culture resources based on big data. Journal of Mathematics, 2022.

[8] Chen, Q. , Tian, Z. , Lei, T. , & Huang, S. . (2023). An association rule mining model for evaluating the potential correlation of construction cross operation risk. Engineering construction & architectural management, 30(10), 5109-5132.

[9] Baamer, R. M. , Iqbal, A. , Lobo, D. N. , Knaggs, R. D. , Levy, N. A. , & Toh, L. S. . (2022). Utility or unidimensional and functional pain assessment toots in adult postoperative patients: a systematic review. British journal of anaesthesia(5), 128.

[10] Lieberman, L. J. , Ball, L. , Beach, P. , & Perreault, M. . (2022). A qualitative inquiry of a three-month virtual practicum program on youth with visual impairments and their coaches. International journal of environmental research and public health, 19(2).

[11] Yang, J. J. , Lo, H. W. , Chao, C. S. , Shen, C. C. , & Yang, C. C. . (2020). Establishing a sustainable sports tourism evaluation framework with a hybrid multi-criteria decision-making model to explore potential sports tourism attractions in taiwan. Sustainability, 12(4), 1673.

[12] Seto, Y. , & Ohtsuka, M. . (2022). Recipro: free and open-source multipurpose crystallographic software integrating a crystal model database and viewer, diffraction and microscopy simulators, and diffraction data analysis tools. Journal of Applied Crystallography(2), 55.

[13] Silveira, A. C. D. , Rodrigues, E. C. , Saleme, E. B. , Covaci, A. , Ghinea, G. , & Santos, C. A. S. . (2023). Thermal and wind devices for multisensory human-computer interaction: an overview. Multimedia Tools and Applications, 82(22), 34485-34512.

[14] Tan, J. , & He, J. . (2024). Thermal radiation image recognition camera using target detection techniques with human computer interaction. Journal of Radiation Research and Applied Sciences, 17(3).

[15] Yu, D. , Zhang, A. , & Gao, Z. . (2023). Fault diagnosis using redundant data in analog circuits via slime module algorithm for support vector machine. Journal of Ambient Intelligence and Humanized Computing, 14(10), 14261-14276.

[16] Mahsa, H. , Abbas, M. , & Reza, G. . (2023). A novel scheme for mapping of mvt-type pb–zn prospectivity: lightgbm, a highly efficient gradient boosting decision tree machine learning algorithm. Natural resources research, 32(6), 2417-2438.

[17] Guo, F. Y. , Zhou, J. J. , Ruan, Z. Y. , Zhang, J. , & Qi, L. . (2022). Hub-collision avoidance and leaf-node options algorithm for fractal dimension and renormalization of complex networks. Chaos: An Interdisciplinary Journal of Nonlinear Science, 32(12), -.

[18] Ishida, S. , Isozaki, M. , Fujiwara, Y. , Takei, N. , Kanamoto, M. , & Kimura, H. , et al. (2023). Effects of the training data condition on arterial spin labeling parameter estimation using a simulation-based supervised deep neural network. Journal of Computer Assisted Tomography, 48(3), 459-471.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 6, 2024
Accepted: Aug 3, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1403–1411, DOI 10.12694/scpe.v26i3.4111

IMPROVEMENT AND OPTIMIZATION OF MACHINE LEARNING ALGORITHMS BASED ON INTELLIGENT COMPUTING

JINDI FU[*], XIAOJIE LIU[†], PENGKAI MA[‡], AND CHANGXIN SONG[§]

Abstract. In this paper, a sensor cloud data intrusion detection framework is proposed. The framework uses parallel discrete optimization techniques for feature refining and incorporates machine learning principles to improve sensing cloud security. Firstly, a set of optimal feature evaluation criteria is established, and a parallel discrete optimization feature extraction system is built to reduce the data dimension and strengthen the stability of feature processing. Then, a widely used discrete optimization algorithm is developed, proving its global convergence. The optimal feature set is obtained through parallel screening feature subsets. Finally, using these features and distributed fuzzy cluster analysis, the intrusion behavior of the sensing cloud is accurately detected. This method incorporates the concept of intelligent iterative evolution and self-regulating clustering strategy, which not only overcomes the local optimal trap that the conventional fuzzy clustering algorithm may encounter but also realizes the automatic adjustment of the number of clusters. The experimental data show that the intrusion detection algorithm performs excellently in providing accurate intrusion determination results. Compared with other detection algorithms, the accuracy of anomaly detection and the reduction of missing detection rate is significantly improved. In addition, the algorithm shows anti-interference solid ability and can maintain stable performance in noisy environments.

Key words: Intelligent computing; Sensing cloud data; Discrete optimization algorithm; Machine learning; Algorithm improvement; Intrusion detection.

1. Introduction.

The purpose of intrusion detection is to detect all kinds of attack intentions quickly and accurately and to respond to them quickly. Abuse recognition and anomaly detection are two standard attack methods. Anomaly detection using machine learning technology can effectively identify potential attacks, so it has received more and more attention. Given the massive perception networks existing in large-scale and high-dimensional perception networks, feature extraction technology is used in the literature [1] to extract highly discriminative feature subsets from the perception networks. Scholars maximize the recognition accuracy to reduce the complexity of the problem. In literature [2], the Markov chain method extracts feature subsets efficiently by combining the maximum information coefficient with the symmetric uncertainty criterion. Literature [3] proposes a criterion based on information increment. Irrelevant and redundant features are gradually eliminated, and the target classification is realized. The above feature extraction method needs to be completed in steps, so it is challenging to meet the real-time processing requirements in the processing process. So, scholars began to use neural networks, rough sets, support vector machines, cluster analysis and other machine learning features to realize the recognition of intrusion behavior. In the literature [4], the FCM algorithm is applied to sensor networks, and an improved fuzzy clustering method, AGFCM, is proposed accordingly. Literature [5] proposes a multi-attribute fusion model based on a genetic algorithm. Literature [6] introduces the learning technique of eliciting C-means to improve its ability to identify sample sets correctly. However, it is easily affected by the initial cluster center, local extreme value, cluster number presetting, etc. Literature [7] established a graphical signal model according to the directional characteristics of each sensor. The smoothness ratio of the image is obtained by processing the image with a low-pass filter. It evaluates the anomalies by statistical inspection of the data in the network and combined with the decision threshold. However, this algorithm needs to train many samples, which significantly disadvantages improving the algorithm's operation speed. Literature [8] uses K-Means to cluster data in the network and uses the K-nearest neighbor algorithm to transmit standard

[*] Keyi College, Zhejiang Sci-Tech University, Shaoxing 312369, China
[†] Shanghai Urban Construction Vocational College, Shanghai 201415, China
[‡] Hefei University of Technology, Xuancheng 242000, China
[§] Shanghai Urban Construction Vocational College, Shanghai 201415, China (Corresponding author, songcx321@163.com)

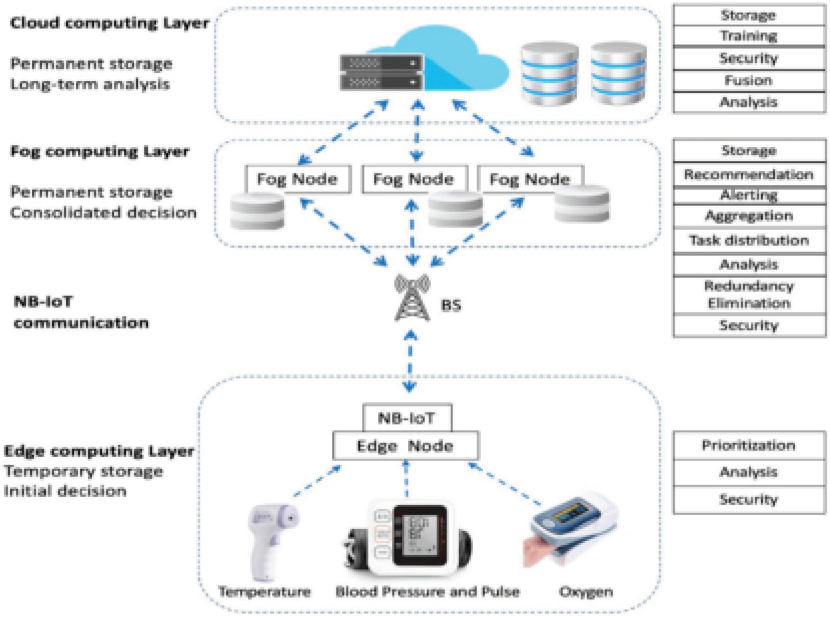


Fig. 2.1: Sensor Cloud architecture based on fog computing.

clustering information to lower-level nodes to partition abnormal data. However, this algorithm focuses on the study of spatial correlation of data and does not effectively combine the statistical characteristics of traffic with it. This results in low accuracy of anomaly recognition. This paper establishes a new perceptual cloud cluster model by combining distributed fuzzy clustering and parallel discrete optimal methods to achieve effective and reliable intrusion detection.

2. Perception-oriented fog source cloud system structure.

A cloud-sensing architecture based on fog computers is established. Cloud computing has enormous computing functions, which can conduct data mining and analysis for massive data in massive perception networks. This provides a reference for the behavior of various users [9]. The "fog" technology is used to realize network computing extending from "cloud" to "edge." The system comprises multiple mobile nodes with certain computing functions to form a virtual network. It can process, store and manage data autonomously or assist the cloud. Figure 2.1 is a schematic of a perceptive cloud architecture based on fog computing.

2.1. Optimal feature extraction method based on parallel discretization.

2.1.1. Extracting feature subsets. There are n samples $\{u_i\}_{i=1}^{n}$ in the data set S, and each sample u_i can be interpreted by m samples of $G_i = (g_{i1}, g_{i1}, \cdots, g_{im})$. And divide the data set S into z categories $Z = \{z_j\}_{j=1}^{z}$. The goal of feature extraction is to extract a subspace composed of t features from m feature sets, so that it has similar recognition performance to the original data. The feature extraction vector P is:

$$\begin{aligned} &P = (p_1, \cdots, p_j, \cdots, p_m) \\ &p_j \in \{0,1\}, P^T 1 = t \end{aligned} \tag{2.1}$$

$p_j = 1$ means that the corresponding feature is extracted. If not, $p_j = 0$. $G_i P^T = \sum_{j=1}^{t} \hat{g}_{ij}$ is obtained after the feature of vector P of u_i is extracted, where E is the corresponding characteristic description after the feature of u_i is extracted [10]. The feature extraction matrix $(\hat{g}_{i1}, \hat{g}_{i2}, \cdots, \hat{g}_{it})$ is determined. There is a corresponding relationship between the extraction matrix E and P, and the element containing only t lines in the extraction

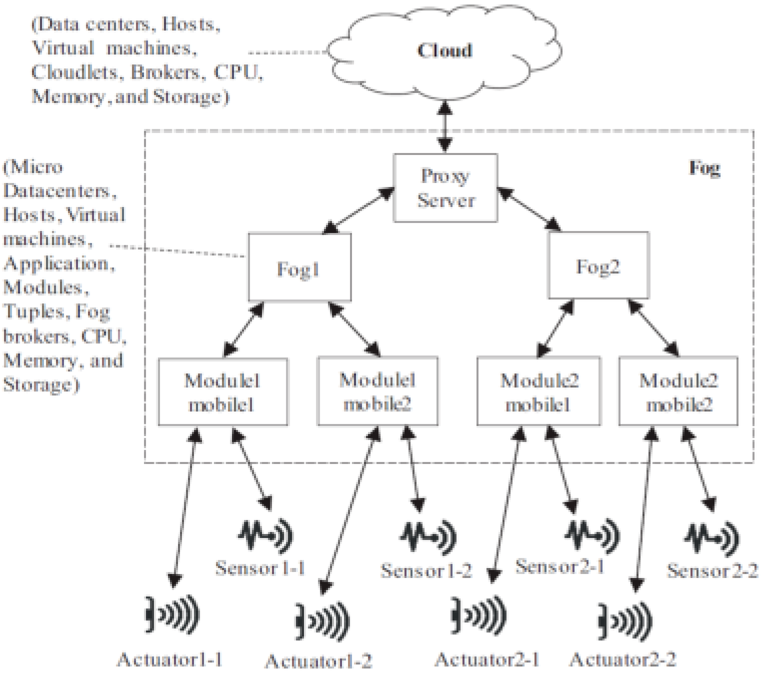

(Data centers, Hosts, Virtual machines, Cloudlets, Brokers, CPU, Memory, and Storage)
Cloud
(Micro Datacenters, Hosts, Virtual machines, Application, Modules, Tuples, Fog brokers, CPU, Memory, and Storage)
Fog
Proxy Server
Fog1
Fog2
Module1 mobile1
Module1 mobile2
Module2 mobile1
Module2 mobile2
Sensor1-1
Sensor1-2
Sensor2-1
Sensor2-2
Actuator1-1
Actuator1-2
Actuator2-1
Actuator2-2

Fig. 2.2: Schematic diagram of feature subset extraction.

matrix E is set to 1 :

$$E_{m\times\Sigma} = \left(P^T, P^T, \cdots, P^T\right) = \begin{bmatrix} p_1 & p_1 & \cdots & p_1 \\ p_2 & p_2 & \cdots & p_2 \\ \vdots & \vdots & & \vdots \\ p_m & p_m & \cdots & p_m \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \|E\|_{2,0} = t \tag{2.2}$$

The eigensubset $\hat{G} = \{\hat{g}_1, \hat{g}_2, \cdots, \hat{g}_t\}$ is extracted by the feature extraction matrix E. Figure 2.2 depicts a specific graph of the E extraction S feature (image cited in Low-latency and energy-efficient scheduling in fog-based IoT applications). Visualization $\hat{G} = \{\hat{g}_1, \hat{g}_2, \cdots, \hat{g}_t\}$ reflects the non-o elements of vector P corresponding to the characteristic value of the sample, then the characteristic subset of the data sample can be extracted [11]. To better evaluate the feature subset G, the evaluation index $\phi(S)$ of the best feature subset is determined.

The best feature subset evaluation index $\phi(S)$ is:

$$\phi(S) = \min_P \left\| \frac{\varphi^T(\Lambda E)(\Lambda E)^T\varphi}{n^2} - C \right\|_G^2 \tag{2.3}$$

where $C = (c_{ij})_{z\times x}$ is a similar matrix between groups. Λ is a matrix of $n \times m$. Where $\varphi = (\lambda_{ij})_{n\times x}$ is the association coefficient between the attribute and the class. $\lambda_{ij} \in [0, 1]$ represents the correlation between sample u_i and class z_j. Use the method of maximum information to solve the λ_{ij} problem. Variables are expressed in terms of $U = \{u_i, i = 1, 2, \cdots, N\}$ and $V = \{v_i, i = 1, 2, \cdots, N\}$, and defined in terms of mutual information about U and V

$$MI(U, V) = \sum_{u\in U}\sum_{v\in V} p(u, v) \lg \frac{p(u, v)}{p(u)p(v)} \tag{2.4}$$

where $p(u, v)$ is the combination possibility of U and V.$p(u), p(v)$ is a function of the marginal probability. The value range of U and V is divided into two parts: c and d. The combined space of U and V is divided into

$c \times d$ grids, and then the histogram is used to estimate $p(u), p(v)$, and the estimate $MI(U, V)_{c,d}$ of $MI(U, V)$ is obtained. According to the different types of class $c \times d$ grid, the maximum mutual information $MI(U, V)_{c,d}^{\max}$ is introduced in Class $c \times d$ grid division. The maximum information factor is defined as:

$$MI(U, V) = \max_{c \times d \leq D(N)} \left\{ \frac{MI(U, V)_{c,d}^{\max}}{\log(c, d)} \right\} \tag{2.5}$$

$D(N)$ represents the maximum value of the grid, which is generally represented by $D(N) = N^{0.6}$. Select $\lambda_{ij} = MIC\left(u_i, z_j\right)$ when determining the more significant amount of information.

Proof that all the features in $\Lambda_{n \times m}$ are composed of a standardized central program, namely $\sum_{i=1}^{n} g_{ij} = 0, \sum_{i=1}^{n} g^2 ij = 1$. If $C' = n^2 C, D = \varphi^T (\Lambda E)(\Lambda E)^T \varphi$, then:

$$\begin{aligned} & \phi(S) = \min_P \frac{1}{n^2} \left\| D - C' \right\|_G^2 = \\ & \min_P \frac{1}{n^2} \operatorname{tr}\left[\left(D - C'\right)(D - C)\right] \Rightarrow \\ & \phi(S) = \min_P \frac{1}{n^2} \operatorname{tr}\left(D^T D + C'^T C' - 2C' D\right) \end{aligned} \tag{2.6}$$

Since $C'TC'$ is a constant matrix, the minimum of $\phi(S)$ requires both $\min_P \operatorname{tr}\left(D^T D\right)$ and $\min_P \operatorname{tr}\left(C^T D\right) \cdot$ $\min_P \operatorname{tr}\left(D^T D\right)$ is represented as follows:

$$\begin{aligned} & \min_P \operatorname{tr}\left(D^T D\right) = \sum_{i,j=1}^{t} \left[\left(\hat{g}_i^T \varphi\right)\left(\hat{g}_j^T \varphi\right)\right]^2 = \\ & \sum_{i,j=1}^{t} \left(\hat{g}_i^T \left(\varphi \varphi^T\right) \hat{g}_j\right) \Rightarrow \min_P \operatorname{tr}\left(D^T D\right) = \\ & \sum_{i,j=1}^{t} \sum_{h}^{z} \left(\left\langle \hat{g}_i, R_h \right\rangle \times \left\langle \hat{g}_j, R_h \right\rangle\right)^2 \Rightarrow \\ & \min_P \operatorname{tr}\left(D^T D\right) = \sum_{i,j=1}^{t} \sum_{h}^{=} n^4 \delta_{R_h}^4 \xi_{g_i, R_h}^2 \xi_{\hat{g}_j, R_h}^2 \end{aligned} \tag{2.7}$$

$R_h (h = 1, 2, \cdots, z)$ is the element of the h line corresponding to the matrix $\varphi_{nx} =$ of class $z_h; \delta_{R_h}^2$ is the standard deviation of $z_h; \xi_{\hat{g}_i, R_h}\left(\xi_{\hat{g}_j, R_h}\right)$ is the Pearson correlation between $\hat{g}_i \left(\hat{g}_j\right)$ and z_h. $\sum_h^z \xi_{\hat{g}_i, R_h}^2 \xi_{\hat{g}_j, R_h}^2$ reflects the redundancy of the corresponding characteristics of $\hat{g}_i$ and $\hat{g}_j$, and the minimum of $\min_P \sum_h^z \xi_{\hat{g}_i, R_h}^2 \xi_{\hat{g}_j, R_h}^2$ is required for $\min_P \operatorname{tr}\left(D^T D\right) = \sum_{i,j=1}^t \sum_h^z n^4 \delta_{R_k}^4 \xi_{\hat{g}_i, R_i}^2 \xi_{\hat{g}_j, R_h}^2$ to be optimal. That is, each element has the least redundancy. Where $\max_P \operatorname{tr}\left(C'T\right)$ is:

$$\begin{aligned} & \max_p \operatorname{tr}\left(C^\tau D\right) = (\Lambda E)^T \varphi C' \phi^T (\Lambda E) = \\ & \sum_{i=1}^{t} \hat{g}_i^T \left(\varphi C' \varphi^\tau\right) \hat{g}_i \Rightarrow \max_p t^r \left(C^{-T} D\right) = \sum_{i=1}^{t} \hat{g}_i^T \left(\sum_{i=1}^{2} \sum_{i=1}^{2} R_e c_o R_i^T\right) \hat{g}_i \end{aligned} \tag{2.8}$$

$\sum_{i=1}^{2} \sum_{l=1}^{2} R_e c_{er} R_i^T$ fully embodies the similarity between attributes and the correlation between attributes and categories. This preserves the original category associations to a maximum extent. The feature extraction vector P and matrix E correspond [12]. After all the positions of $p_j = 1$ in P are determined, a specific representation of E can be obtained. Then, feature extraction is carried out.

2.1.2. Construct the model of feature subset extraction. The iterative, evolutionary algorithm is adopted to solve the optimal problem by imitating the life activity of organisms or the physical property change of materials. This project proposes a discrete optimization method suitable for the existing intelligent optimization methods [13]. It is used to solve $\phi(S)$ to obtain the optimal feature extraction vector P. This project intends to use the parallel algorithm and Q-value discrete optimal method to classify $\mathrm{Q} - P_i, i = 1, 2, \cdots, Q$ multidimensional data to obtain more robust and better classification effect. This can improve the robustness and reliability of the algorithm. This paper adopts P_{bost} language and distributed fuzzy clustering method. An initial particle population $H = \{U_i\}_{i=1}^{N}$ with N scales is randomly generated in the solution space L^m. Each particle $U_i = (u_{nl}, u_{i2}, \cdots, u_{im})$ is a possible solution that evolves repeatedly until it reaches an optimal solution.

In this discrete optimization algorithm, the correction method for particle $U_i(t)$ is as follows:

$$U_i(t+1) = \begin{cases} U_i(t) + \Delta, l_1 \geq \kappa \\ U_i(t), l_1 < \kappa \end{cases}$$
$$\Delta = \begin{cases} \eta_1 \otimes (U_i(t) \leftrightarrow U_i(t)), l_2 \leq \alpha_1 \\ \eta_2 \otimes (U_i(t) \leftrightarrow U_{b=2}(t)) + l_2 \otimes \\ (U_i(t) \leftrightarrow U_8(t)), \alpha_1 < l_2 \leq \alpha_2 \\ \eta_3 \otimes (U_i(t) \leftrightarrow U_j(t)), \alpha_2 < l_2 < 1, i \neq j \end{cases} \tag{2.9}$$

where $\kappa, \alpha_1, \alpha_2$ is the Update control probability. l_1, l_2 is any random number in the interval $(0,1); C \leftrightarrow D$ stands for individual C research on D; Where $\eta_1 \otimes (\leftrightarrow)$ represents the size of the degree of learning, and the larger the value of η_1, the more significant the role of particle D on the evolution direction of C. To improve the algorithm's convergence, this project plans to divide H into several subgroups $H_j, j = 1, 2, \cdots, O$. Each subpopulation adaptively sets the size of α_1, α_2 and η_1, η_2, η_3 according to the fitness of particles, and then automatically adjusts the size of particle swarm to realize the real-time adjustment of learning objectives and learning intensity of particles [14]. About the existence of subgroups H_j :

$$\begin{cases} \alpha_1(\eta_1) \propto \ln\left(\frac{\max g(U_i(t))}{g(U_{k=\alpha}(t))} \frac{t}{T_{\max}} + 1\right) \\ \alpha_2(\eta_2, \eta_3) \propto \ln\left(2 - \frac{\max_{U_e H_i} g(U_i(t))}{g(U_{\text{bai}}\ (t))} \frac{t}{T_{\max}}\right) \end{cases} \tag{2.10}$$

where $g(\cdot)$ is the objective function value. $T_{\max}$ is the maximum number of iterations. For H_j with better fitness, the particle has a higher self-learning ability. That is, it evolves through its mutations, thus effectively expanding the deep optimization space of the algorithm. For H_j with poor fitness, the particle will tend to the optimal value of the population and the historical optimal value with a greater probability of accelerating the evolution.

2.2. Implementation of sensing cloud intrusion detection. Test case u_i^{bul} selects cluster centers closer to itself as categories, and obtains Z sample categories $S_z^{(\infty)}, z = 1, 2, \cdots, Z$ after all sample sets are classified [15]. Because there is usually very little data for this to happen, it is very vulnerable when there is only a small amount of data S_z^{not} . To further determine the data anomalies, the categories of suspected $S_=^{\text{sax}}$ and ordinary correction data sets are compared sequentially. If the following conditions are met:

$$\min_{H \to Z} \left| \frac{\sum_= (u_i^{vax} - v_s^{vax})}{|S_=^{\text{lom}}|} - \frac{\sum_{H \in SS_N} (u_j - v_H)}{|S_H|} \leq \theta \right| \tag{2.11}$$

Then it is inevitable that something abnormal has occurred in S_z^{text} . Where Z' is the number of classes in the standard correction data set; v_z^{text} , v_H is the cluster center of detection data classification S_z^{text} and normal data classification S_H respectively. $|S_H|$ represents the internal data size of the class.

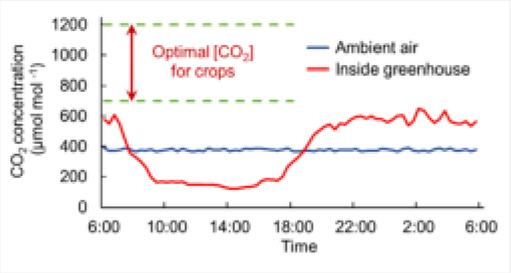


Fig. 3.1: Changes in carbon dioxide concentration values.

3. Simulation experiment.

3.1. Experimental environment. A greenhouse is taken as an example for experimental study. Conventional multi-station environmental parameter measurement system adopts cable transmission mode, which requires cable transmission when transmitting data, which increases the cost of system design, installation and later maintenance [16]. Combine discrete optimization algorithms with machine learning to optimize the wiring of communication and power supply to improve the facility's operational efficiency. The system has important theoretical and practical significance. A wireless sensor network of CO$_2$ concentration in the greenhouse was established using a discrete optimization algorithm and machine learning algorithm to realize online adjustment of greenhouse gas concentration. The following is the dynamic function model of wireless sensor network monitoring system:

$$x_{CO_1} = RC - B(u) \tag{3.1}$$

RC is the original CO$_2$ content in the greenhouse and $B(u)$ is the direct effect of photosynthetic efficiency of crops in the greenhouse on soil carbon content. The content of CO$_2$ showed a dynamic change with the experiment. The content of CO$_2$ was detected within 2 hours, and the corresponding results were obtained. Figure 3.1 shows the change in C content in the greenhouse. The greenhouse covers an area of 300 m *1000 m and is evenly arranged with 30x100 sensors in different directions. Then MATLAB is used as the test platform. Image signal processing in reference [1] is compared with hierarchical clustering in reference [2].

3.2. Analysis of experimental results. The detection rate refers to the probability of accurately detecting an anomaly, while the false alarm refers to the probability of being detected. A good anomaly detection algorithm should have high and small false favorable rates to ensure the final result's accuracy [17]. The discreteness of a node is the probability of finding an anomaly when a node is in a good state. This project takes the abnormal ratio of nodes in the network as the test target. The abnormal rate ranges from 5% to 25%. The effect of the anomaly ratio of the three algorithms on anomaly detection is shown in FIG. 3.2 and FIG. 3.3.

It can be seen from Figure 3.2 and Figure 3.3 that when the number of detection rates is 25, the method of the invention can obtain a detection rate of 94.48%. The detection rate of graph signal processing was 89.17%. The stratified polymerization method was 84.58%. The detection rate of the three algorithms decreases gradually with the increase of the node dispersion rate, while the false alarm increases continuously [18]. However, the accuracy of the proposed algorithm is significantly higher than that of graph processing and hierarchical clustering because the algorithm uses the discrete optimization algorithm in the cloud computing environment to remove excess information in the data and thus improve the accuracy of anomaly detection. Next, the amount of node energy lost when three methods are used to detect data anomalies is analyzed. As shown in Figure 3.4, node power consumption gradually increases with test times [19]. The proposed algorithm requires the lowest power consumption, followed by hierarchical clustering, and the last is graph signal processing. It is proved that the algorithm proposed in this paper can complete real-time monitoring of sensing data with the lowest power consumption.

Figure 3.5 compares the data anomaly detection time of the three methods. The number of tests is 300, and the comparison of abnormal data detected by the three methods is shown in Figure 3.5 as the average time

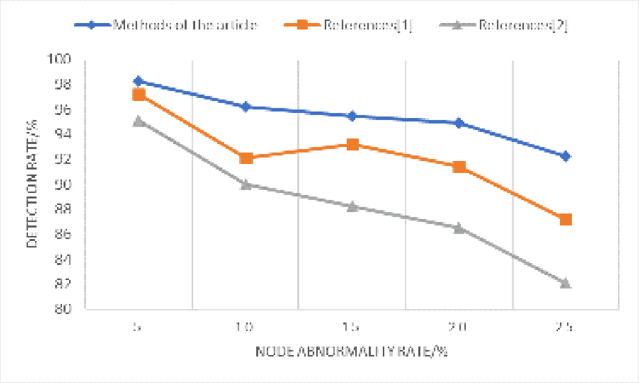


Fig. 3.2: Comparison of data anomaly calculation detection rates.

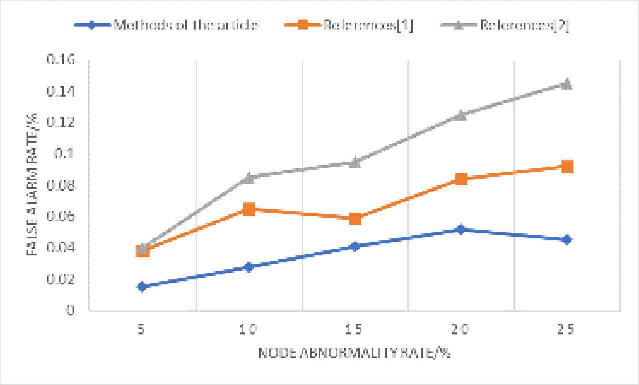


Fig. 3.3: Comparison of false alarm rates of data anomaly detection.

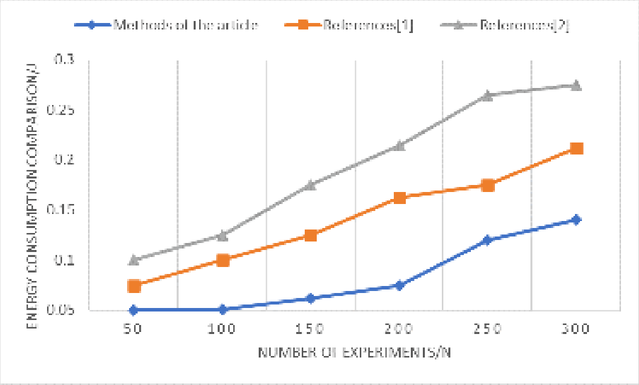


Fig. 3.4: Comparison of energy consumption of abnormal detection nodes.

of each cycle. The algorithm in this paper does not change the anomaly detection time of data and has good stability. Its detection time is minimal and it has a good application prospect.

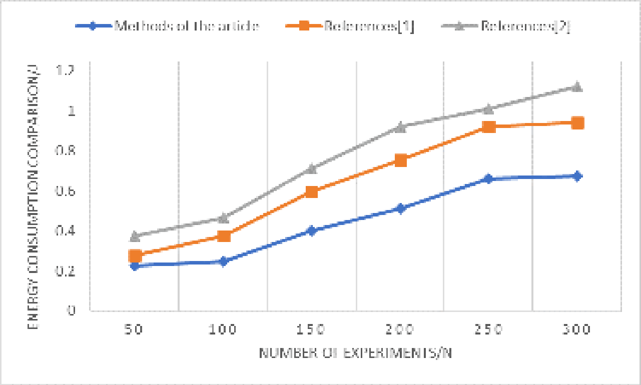


Fig. 3.5: Comparison of data anomaly detection time.

4. Conclusion. In this paper, an innovative sensor cloud intrusion detection algorithm is proposed, which cleverly combines the characteristics of discrete optimization algorithm (DOA) and machine learning technology and effectively copes with the challenges of large data scale, high dimension and variable intrusion behavior in the sensor cloud environment. The optimal feature combination is extracted by parallel feature subset screening, which further enhances the detection ability of the algorithm. Finally, using distributed fuzzy clustering technology and intelligent iterative evolution ideas, the algorithm realizes the automatic partition of cluster numbers while avoiding local optimization, thus significantly improving the efficiency and accuracy of intrusion detection. By defining the optimal feature evaluation index and constructing the parallel discrete optimization feature extraction framework, this paper successfully reduces the data dimension and improves the robustness of feature extraction, thus laying a solid foundation for subsequent intrusion detection.

REFERENCES

[1] Shen, Y., Liu, Y., Tian, Y., & Na, X. (2022). Parallel sensing in metaverses: Virtual-real interactive smart systems for ”6S” sensing. IEEE/CAA Journal of Automatica Sinica, 9(12), 2047-2054.

[2] Lu, H., Zong, Q., Lai, S., Tian, B., & Xie, L. (2021). Flight with limited field of view: A parallel and gradient-free strategy for micro aerial vehicle. IEEE Transactions on Industrial Electronics, 69(9), 9258-9267.

[3] Cong, P., Zhou, J., Chen, M., & Wei, T. (2020). Personality-guided cloud pricing via reinforcement learning. IEEE Transactions on Cloud Computing, 10(2), 925-943.

[4] Balaji, K., Sai Kiran, P., & Sunil Kumar, M. (2023). Power aware virtual machine placement in IaaS cloud using discrete firefly algorithm. Applied Nanoscience, 13(3), 2003-2011.

[5] Srivastava, A., & Kumar, N. (2023). Multi-objective binary whale optimization-based virtual machine allocation in cloud environments. International Journal of Swarm Intelligence Research (IJSIR), 14(1), 1-23.

[6] Cong, P., Zhang, Z., Zhou, J., Liu, X., Liu, Y., & Wei, T. (2021). Customer adaptive resource provisioning for long-term cloud profit maximization under constrained budget. IEEE Transactions on Parallel and Distributed Systems, 33(6), 1373-1392.

[7] Dong, W., Lao, Y., Kaess, M., & Koltun, V. (2022). ASH: A modern framework for parallel spatial hashing in 3D perception. IEEE transactions on pattern analysis and machine intelligence, 45(5), 5417-5435.

[8] Naghdehforoushha, M., Fooladi, M. D. T., Rezvani, M. H., & Sadeghi, M. M. G. (2022). BLMDP: A new bi-level Markov decision process approach to joint bidding andtask-scheduling in cloud spot market. Turkish Journal of Electrical Engineering and Computer Sciences, 30(4), 1419-1438.

[9] Yang, H., & Carlone, L. (2022). Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization. IEEE transactions on pattern analysis and machine intelligence, 45(3), 2816-2834.

[10] Wang, X., Han, S., Yang, L., Yao, T., & Li, L. (2020). Parallel internet of vehicles: ACP-based system architecture and behavioral modeling. IEEE Internet of Things Journal, 7(5), 3735-3746.

[11] Mishra, N., & Singh, R. K. (2020). DDoS vulnerabilities analysis and mitigation model in cloud computing. Journal of Discrete Mathematical Sciences and Cryptography, 23(2), 535-545.

[12] Cong, P., Xu, G., Wei, T., & Li, K. (2020). A survey of profit optimization techniques for cloud providers. ACM Computing Surveys (CSUR), 53(2), 1-35.

[13] Zhou, B., Pan, J., Gao, F., & Shen, S. (2021). Raptor: Robust and perception-aware trajectory replanning for quadrotor fast flight. IEEE Transactions on Robotics, 37(6), 1992-2009.

[14] Zhang, Y., Zhou, Y., Lu, H., & Fujita, H. (2020). Traffic network flow prediction using parallel training for deep convolutional neural networks on spark cloud. IEEE Transactions on Industrial Informatics, 16(12), 7369-7380.

[15] Chakraborty, S., Saha, A. K., & Chhabra, A. (2023). Improving whale optimization algorithm with elite strategy and its application to engineering-design and cloud task scheduling problems. Cognitive Computation, 15(5), 1497-1525.

[16] Tang, Q., Fei, Z., Li, B., & Han, Z. (2021). Computation offloading in LEO satellite networks with hybrid cloud and edge computing. IEEE Internet of Things Journal, 8(11), 9164-9176.

[17] Chhabra, A., Singh, G., & Kahlon, K. S. (2021). Multi-criteria HPC task scheduling on IaaS cloud infrastructures using meta-heuristics. Cluster Computing, 24(2), 885-918.

[18] Gualtieri, M., & Platt, R. (2021). Robotic pick-and-place with uncertain object instance segmentation and shape completion. IEEE robotics and automation letters, 6(2), 1753-1760.

[19] Swathi, V. N. V. L. S., Kumar, G. S., & Vathsala, A. V. (2023). Cloud Service Selection System Approach based on QoS Model: A Systematic Review. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 05-13.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 8, 2024
Accepted: Jul 15, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1412–1421, DOI 10.12694/scpe.v26i3.4127

WHALE OPTIMIZATION ALGORITHM FOR EFFICIENT TASK ALLOCATION IN THE INTERNET OF THINGS

WANCHANG SHU*

Abstract. In order to solve the problem of reducing worker costs and improving worker efficiency, the author proposes a title whale optimization algorithm for efficient task allocation in the Internet of Things. This algorithm adopts the fuzzy chance constrained programming method to model the online time of workers, and introduces delay costs and idle costs based on whether there is delay or not, Due to the fact that the corresponding problem is a combinatorial optimization problem and belongs to the NP hard problem category, a two-stage task allocation algorithm is designed to solve it in combination with the whale optimization algorithm. The experimental results show that after being simulated by the algorithm, half of the workers reached the highest efficiency of 1, and the expected online time of the workers was less than 30, and the task execution time of the workers was less than 35. The task allocation algorithm designed by the author has higher worker efficiency compared to other algorithms and has broad application prospects.

Key words: Group intelligence perception, Task allocation, Worker costs, Worker efficiency, Whale Optimization Algorithm

1. Introduction. The perception layer of the Internet of Things system is composed of a large number of heterogeneous devices, and devices with different functions collaborate to complete tasks generated in the Internet of Things environment, enabling the system to quickly respond to user requests and achieve intelligent services [1]. Due to limited device resources and computing power, devices can only autonomously execute tasks that meet their capabilities and resources in a dynamic environment. Otherwise, it will lead to imbalanced load on terminal devices and system instability, thereby reducing user experience [2]. Therefore, task allocation is crucial, that is, how to reasonably allocate IoT tasks to terminal devices and meet the limited computing power and resources of terminal devices.

The task publisher is the user who uploads the task to the group intelligence perception platform. The group intelligence perception server is a server that performs task processing, allocation, and data processing. Workers are the group of people who perform tasks through the mobile terminals they carry [3]. After the task publisher in the group intelligence perception system uploads the task and its requirements to the cloud server, the group intelligence perception server will preprocess the task, such as large-scale task decomposition, similar task fusion, etc. [4]. If workers carrying mobile terminals are interested in certain tasks in the system, they will register and provide their own information on the server, and then the server selects suitable workers as executors to perform related tasks. The selected workers use rich sensors to perceive various data and perform different types of tasks [5]. After the workers complete the task, the swarm intelligence perception server collects the task data provided by the workers and processes it, and then returns the results to the task publisher [6]. In the swarm intelligence perception system, there are multiple workers waiting to execute tasks at the same time, and the server needs to assign multiple tasks to the workers in order to better complete the tasks and provide high-quality data. It is crucial to efficiently allocate tasks to workers, and accordingly, this type of task allocation problem has become a research hotspot in the field of swarm intelligence perception [7].

2. Literature Review. Crowd sensing (CS) utilizes intelligent devices to collect data and provide large-scale application services that static sensor networks cannot support [8]. In order to better support various application services, swarm intelligence perception systems require tasks to be assigned to suitable workers under certain constraints, and workers move to corresponding positions to perform tasks. This task allocation problem is currently a hot topic in the research field of swarm intelligence perception [9]. Currently, scholars have conducted research on task allocation in swarm intelligence perception systems. Mahmoud, M. M. et

*Jiangxi Teachers College, Yingtan, Jiangxi, China, 335000 (Corresponding author, wanchang_shu93@163.com)

al. applied the whale optimization algorithm based on fractional order proportional integral controller to the unified power quality regulator and STATCOM tool. They operate best with the help of improved control systems to improve system reliability and fast dynamic response, and reduce total harmonic distortion, thereby improving power quality [10]. P. C. H. et al. studied a unique multi-objective whale optimization (MOWOA) algorithm for solving multi-objective problems. In order to verify its effectiveness, the author applied the proposed method to the IEEE-33 and IEEE-69 radial bus distribution systems. It was found that the proposed method can improve power loss, reduce annual economic losses, and improve voltage distribution [11]. Wen et al. proposed an efficient index based multi-objective evolutionary algorithm with a mixed encoding scheme of task execution order and robot starting point information. This algorithm uses supercapacity indicators for environment selection to enhance convergence, and uses modified crowding distances for file updates to promote diversity [12]. In order to this end, considering the flexible online time of workers, a fuzzy chance constrained programming method is adopted to model their online time, and delay costs and idle costs are introduced. The corresponding task allocation problem is a combinatorial optimization problem, belonging to the category of NP hard problems. There is no time efficient optimal algorithm, and only suboptimal algorithms can be considered [13]. Given the strong global search capability of Whale Optimization Algorithm (WOA), a two-stage algorithm was designed using WOA to solve the task allocation problem. The simulation results show that the proposed algorithm has better search performance compared to other algorithms; Meanwhile, compared to fixed online time, considering flexible online time results in higher worker efficiency and lower worker costs.

3. Method.

3.1. System Model. Consider a swarm intelligence perception system with t perception tasks and w registered workers. Among them, $T = \{T_1, \cdots, T_t\}$ and $W = \{W_1, \cdots, W_w\}$ represent task sets and worker sets, respectively. For task i, TT_i is the time required to execute the task; For worker j, WT_j is the estimated online time set by the worker during registration. For the convenience of describing the problem, provide the following definitions.

Definition 1. Task Execution Time Mission Time: The task execution time is the total time spent by workers executing tasks assigned by the system, as defined in Equation 3.1

$$MT_j = \sum_{T_q \in V_j} TT_q \tag{3.1}$$

Among them, V_j represents the set of tasks assigned to workers.

Definition 2. Idle Time: The time during which a worker does not perform a task within the expected online time is called idle time, which is defined as Equation 3.2

$$IT_j = WT_j - MT_j \tag{3.2}$$

At this point, $WT_j \geqslant MT_j$.

Definition 3. Delay time: When the worker's task execution time exceeds the worker's expected online time, the actual online time of the worker is the worker's task execution time, and the part of the worker's time exceeding the expected online time is the delay time, which is defined as equation 3.3:

$$DT_j = MT_j - WT_j \tag{3.3}$$

According to the definition of worker idle time, the cost of worker idle time is shown in equation 3.4:

$$IC_j = \alpha * IT_j \tag{3.4}$$

Among them, α is the unit idle time cost.

In addition, according to the definition of worker delay time, the cost of worker delay is shown in equation 3.5:

$$DC_j = \beta * DT_j \tag{3.5}$$

Among them, β represents the delay cost per worker unit.

Before task allocation, as workers have not yet started executing the task, they cannot determine whether they will choose to extend their online time after executing the task. At this time, the system considers that workers will not extend their online time, thus assigning tasks to them for the first time. $V' = \{V'_0, V'_1, V'_2, V'_3, \cdots, V'_w\}$ represents the allocation result of the perception task at this time. Among them, V'_0 is the set of unassigned tasks in the perception system; $V'_1 \sim V'_w$ represents the set of tasks assigned by the perception system to workers. At this stage, due to the system's consideration that workers do not extend their online time, their task execution time does not exceed their expected online time. At this point, if workers have idle time, there will be an idle cost IC'_j, and the total cost of workers at this time (Total Cost) is determined to be

$$TC' = \sum_{j=1}^{w} IC'_j \tag{3.6}$$

After the initial task allocation, workers can decide whether to consider extending their online time based on their own availability and subsequent arrangements. At this point, the possibility of worker time constraints (that is the possibility that workers do not choose to extend the time) and the level of confidence are used to indicate whether workers ultimately choose to delay. According to the fuzzy chance constrained programming method, when the possibility of worker time constraint is greater than the confidence level, workers must ensure that the task execution time does not exceed the expected online time. At this time, workers do not choose to extend their online time to perform additional tasks; When the likelihood of worker time constraints is less than the confidence level, workers choose to extend their online time to perform additional tasks, and at this point, additional tasks are assigned to workers [14]. Based on this, adjust the results of the initial task allocation. Set $V = \{V_0, V_1, V_2, V_3, \cdots, V_w\}$ represents the final task allocation result. Similarly, V_0 represents the set of tasks that have not been assigned after the final task allocation; $V_1 \sim V_w$ represents the final set of task assignments for workers. After the allocation is completed, the worker cost is determined as either the delay cost or the idle cost based on whether the worker chooses to delay

$$TC_j = \begin{cases} DC_j, MT_j > WT_j \\ IC_j, WT_j \geqslant MT_j \end{cases} \tag{3.7}$$

The corresponding worker efficiency can also be divided into two situations

$$X_j = \begin{cases} 1, MT_j > WT_j \\ \frac{MT_j}{WT_j}, WT_j \geqslant MT_j \end{cases} \tag{3.8}$$

When workers choose to extend their online time, there is no idle time for them, and their efficiency reaches its maximum value of 1; When workers do not choose to extend their online time, they may have idle time. At this time, worker efficiency is the ratio of worker task execution time to worker expected online time, and the total cost for workers is determined to be $TC = \sum_{j=1}^{w} TC_j$

Based on the above definition, the task allocation model considering worker's flexible online time is given as follows

$$max \sum_{j=1}^{w} X_j \tag{3.9}$$

Constraint condition:

$$\sum_{j=0}^{w} |V_j| = t \tag{3.10}$$

$$V_0 \cup V_1 \cup \cdots \cup V_w = T \tag{3.11}$$

$$V_0 \cap V_j = \varnothing, W_j \in W \tag{3.12}$$

$$V_k \cup V_j = \varnothing, j \neq k \tag{3.13}$$

$$|V_j| \geqslant 1, W_j \in W \tag{3.14}$$

$$Cr(WT_j - MT_j \geqslant 0) > \tau, W_j \in W \tag{3.15}$$

$$TC \leqslant TC' \tag{3.16}$$

$$\exists WT_j, WT_j - TT_i \geqslant 0, T_i \in T, W_j \in W \tag{3.17}$$

Among them, equation 3.9 is the optimization objective, which is to maximize the total efficiency of chemical workers; Equations 3.10 and 3.11 indicate that the assigned task is a task published in the system; Equations 3.12 and 3.13 indicate that a task cannot appear in both the worker task set and the unassigned task set simultaneously; Equation 3.14 indicates that each worker needs to complete at least one task; Equation 3.15 is a fuzzy chance constraint on the online time of workers, among them, τ is the confidence level, which is modeled using a fuzzy chance constrained programming model, allowing workers to exceed their expected online time to a certain extent; Equation 3.16 indicates that the total cost of workers after introducing flexible time cannot exceed the total cost of workers without introducing flexible time; Equation (17) indicates that there is at least one worker whose expected online time exceeds the longest required execution time for the task, ensuring that the task with the longest execution time has a chance to be executed during the initial allocation [15]. Due to the fact that the task allocation problem considering worker online time elasticity is a combinatorial optimization problem, belonging to the category of NP hard problems, there is no time efficient optimal algorithm, and only suboptimal algorithms can be considered. Therefore, intelligent algorithms are considered for solving [16]. Compared to other intelligent algorithms, Whale Optimization Algorithm can better balance the global and local optimization stages and has faster convergence speed. Therefore, when solving the task allocation problem determined by equations 3.9 to 3.17, Whale Optimization Algorithm is considered.

3.2. Whale Optimization Algorithm Process. The idea of whale optimization algorithm originates from the predatory behavior of humpback whales. The algorithm is divided into three stages based on the predatory behavior of humpback whales: Surround prey, bubble net attack, and search for prey. Since the whale optimization algorithm was originally proposed to solve continuous problems, and the above-mentioned task allocation problem is a combinatorial optimization problem, it is necessary to improve the whale optimization algorithm to make it more suitable for solving this task allocation problem [17].

Surrounding prey stage. The humpback whale updates its position based on the prey's position, thus approaching the prey, known as the surrounding prey stage [18]. The relevant definitions are shown in equation 3.18:

$$\overrightarrow{X}(t+1) = \overrightarrow{X}^*(t) - \overrightarrow{A} \cdot \overrightarrow{D} \tag{3.18}$$

Among them, t is the current number of iterations; $\overrightarrow{X}^*$ is the current location of the nearest humpback whale to its prey.The coefficients $\overrightarrow{A}$ and distance $\overrightarrow{D}$ are defined as equations 3.19 and 3.20 , respectively

$$\overrightarrow{A} = 2\overrightarrow{a} \cdot \overrightarrow{r} - \overrightarrow{a} \tag{3.19}$$

$$\overrightarrow{D} = |2\overrightarrow{a} \cdot \overrightarrow{r} - \overrightarrow{a}| \tag{3.20}$$

Among them, the sizes of each element in $\overrightarrow{a}$ are between and linearly decrease with increasing iteration times, while the sizes of each element in $\overrightarrow{r}$ are between; $\overrightarrow{C} = 2 \cdot \overrightarrow{r}$ is the coefficient; $\overrightarrow{X}(t)$ is the current position of the humpback whale.

Bubble net attack stage. During the bubble net attack stage, the humpback whale spirals and spits out bubbles to surround its prey. At this point, the spiral motion trajectory of the humpback whale is defined as equation 3.21

$$\overrightarrow{X}(t+1) = \overrightarrow{D}' \cdot e^{bl} \cdot cos(2nl) + \overrightarrow{X}^*(t) \tag{3.21}$$

Among them, $\overrightarrow{D}' = |\overrightarrow{X}^*(T) - \overrightarrow{X}(T)|$ represents the distance between the position of the humpback whale and the prey position, the constant b defines the range of the spiral motion trajectory of the humpback whale, and l is a random number between [-1,1].

Based on the above two stages, at this point, the humpback whale is in the process of discovering prey and moving towards it. Therefore, these two stages are also known as the local search stage [19]. Through observation, it was found that the swimming behavior of humpback whales around their prey includes both surrounding and bubble net attacks. Therefore, in order to describe the behavior of the humpback whale at this time, assuming that the probability of the humpback whale surrounding the prey and the probability of bubble net attack are each 50%, the behavior of the humpback whale is summarized as follows:

$$\overrightarrow{X}(t+1) = \begin{cases} \overrightarrow{X}^*(t) - \overrightarrow{A} \cdot \overrightarrow{D}, p < 0.5 \\ \overrightarrow{D}' \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X}^*(t), p \geqslant 0.5 \end{cases} \tag{3.22}$$

Among them, p is a random number between [0,1].

Search for prey stage. At this stage, humpback whales are still in the search for prey stage, and they randomly search in space based on each other's positions [20]. Therefore, this section is also known as the global search stage, which is defined as follows:

$$\overrightarrow{D} = |\overrightarrow{C} \cdot \overrightarrow{X}_{rand} - \overrightarrow{X}| \tag{3.23}$$

$$\overrightarrow{X}(t+1) = \overrightarrow{X}_{rand} - \overrightarrow{A} \cdot \overrightarrow{D} \tag{3.24}$$

Among them, $\overrightarrow{X}_{rand}$ is the current position of a random whale.

3.3. Encoding and Improvement. Due to the fact that the Whale Optimization Algorithm was originally proposed to solve continuous optimization problems, and the task allocation of worker flexible online time is a combinatorial optimization problem, which involves task and worker pairing problems, therefore, it is necessary to encode the task and worker sequence, and improve the whale optimization algorithm to solve the task allocation problem.

Encode tasks and workers into two separate sequences. $[N_1, N_2, N_3, \cdots, N_t]$, represents the arrangement of t tasks, where; $N_i \in \{1, 2, 3, \cdots, t\}$;$[m_1, m_2, m_3, \cdots, m_i]$ represents the arrangement of executing workers corresponding to the task, where $m_i \in \{0, 1, 2 \cdots, w\}$. When $m_j = 0$, it indicates that the N_j task in the task sequence has not been assigned; When $m_j \neq 0$ occurs, it indicates that the task at the corresponding position is assigned to the worker at the corresponding position, thereby determining the task allocation result. The corresponding task sequence and worker sequence are combined to form a whale, and the total worker efficiency is the fitness value of the whale optimization algorithm. When using the whale optimization algorithm to solve combinatorial optimization problems, additional inversion modules and local search modules are designed to ensure the search performance of the algorithm. In order to better describe the reversal module and local search module, it is assumed that there are 9 tasks and 3 workers in the system. The initialization task sequence is [123456789], and the randomly generated worker sequence is [123123121]. Tasks at the same position and workers form a task worker pair, indicating that the task is executed by the worker.

Inversion module. If the starting point for inversion is 4 and the inversion length is 4, the task sequence that needs to be reversed is 4567; After reversal, the task sequence becomes [123765489]. Based on the optimized task and worker sequence, it can be observed that the task allocation has changed.

Table 3.1: Parameter Settings

Parameter Name	Parameter Value
Task quantity t	70
Number of workers w	10
Task required time TT_i	[5,15]
Estimated online time for workers WT_j	[10,30]
Unit idle cost α	2
Unit delay cost β	1
confidence level τ	0.5

Local search. Select the 5th element in the task sequence for local search optimization. Therefore, remove the 5th element, that is task 6, and select the 2nd position to reinsert the task. The task sequence becomes [162375489]. Based on the randomly generated worker sequence, it can be seen that the task allocation has changed, causing the optimization results of task allocation to jump out of local optima.

3.4. Simulation analysis. The performance of the designed task allocation algorithm was verified through simulation experiments, and the system parameters are shown in Table 3.1. Due to workers choosing to delay, they will perform additional tasks and gain additional benefits, which reduces the cost of delay for workers; When workers are idle, due to not performing tasks without additional benefits, the idle cost cannot be reduced. Therefore, the unit delay cost for workers is set to be less than the unit idle cost for workers. Based on the parameter settings in Table 3.1, we first analyze the impact of confidence level on the number of workers who choose elastic time in each algorithm. Then, compare and analyze the performance of the proposed task allocation algorithm with those based on genetic algorithm, greedy algorithm, and random allocation when the number of workers changes, confidence level changes, and flexible worker number changes. Among them, in the task allocation algorithm based on genetic algorithm, genetic algorithm is used to optimize the preliminary allocation results; In the task allocation algorithm based on greedy algorithm, workers are sequentially assigned tasks that maximize the overall efficiency of the current workers; In a random task allocation algorithm, tasks are randomly assigned to workers based on whether they choose flexible time. Finally, verify the advantages of considering worker flexible time compared to not considering worker flexible time. This algorithm is based on MATLAB R2014a as the simulation platform, with the machine configuration being Intel®$Core^{TM}i7-$4710MQ 2.50GHz 8GBRAM and the operating system being Windows.

4. Conclusion and Discussion. The author uses the fuzzy chance constraint method to model the online time of workers. Workers can choose whether to delay to perform additional tasks based on their own situation. Therefore, the number of workers who choose flexible time in each simulation may be different, which affects the optimization results. In the simulation, the confidence level is pre-set to randomly generate the time constraint possibility of workers. When the time constraint possibility of a worker is greater than the confidence level, it is judged that the worker does not choose elastic time. When the probability of the worker is less than the confidence level, the worker chooses elastic time. It can be seen that the size of the confidence level will affect the number of workers who choose elastic time. Workers who choose elastic time are called elastic workers. Table 4.1 presents the average number of flexible workers in task allocation algorithms based on whale optimization algorithm, genetic algorithm, random allocation, and greedy algorithm at different confidence levels after repeated experiments. From Table 4.1, it can be seen that as the confidence level increases, the number of flexible workers in the algorithm also increases, and the proportion of flexible workers in the total number of workers is approximately equal to the confidence level. This indicates that the number of workers who choose flexible time is influenced by the confidence level. The higher the confidence level, the more workers can meet the delay requirements, and at this point, more workers choose flexible time.

In order to verify the impact of confidence levels on worker efficiency, multiple simulations were conducted at confidence levels of [0.3, 0.4, 0.5, 0.6, 0.7] to obtain the results in Figure 4.1. It can be observed that at different confidence levels, compared to genetic algorithms, random allocation algorithms, and greedy algorithms, whale optimization algorithms achieve the highest worker efficiency. In addition, worker efficiency increases with the

Table 4.1: Changes in the number of flexible workers with confidence levels

Algorithm	Confidence level				
	0.3	0.4	0.5	0.6	0.7
WOA	2.65	3.78	4.73	5.81	6.65
Genetic	2.73	3.85	5.21	6.07	6.73
RA	2.64	3.88	4.76	5.71	6.77
Greedy	3.10	4.02	5.10	6.03	7.12

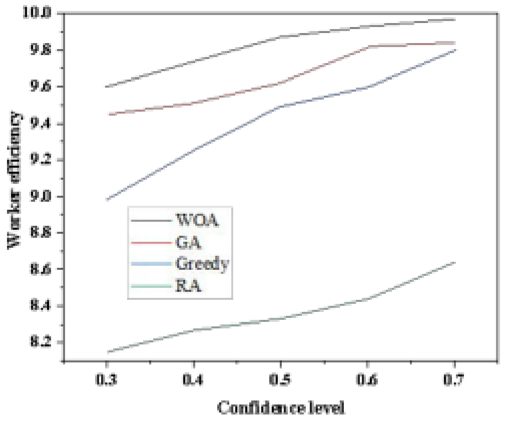


Fig. 4.1: Worker efficiency changes with confidence level

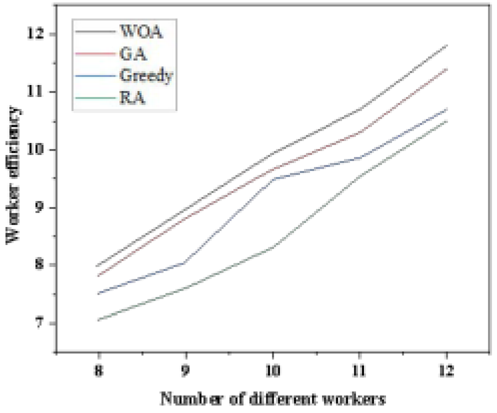


Fig. 4.2: Worker efficiency changes with the number of workers

increase of confidence level in different algorithms.

Figure 4.2 shows the trend of worker efficiency changing with the number of workers. It can be seen that as the number of workers increases, the overall efficiency of workers also increases accordingly, and the task allocation algorithm based on whale optimization algorithm always achieves better worker efficiency than other

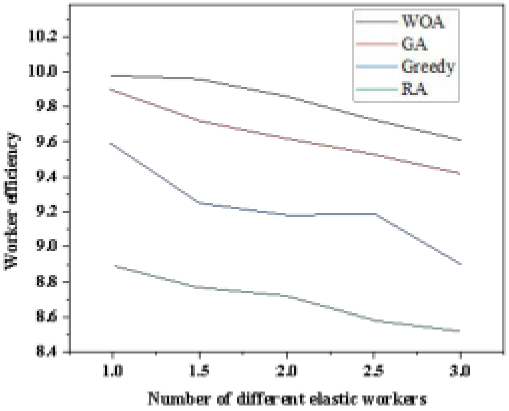


Fig. 4.3: Worker efficiency changes with the number of flexible workers

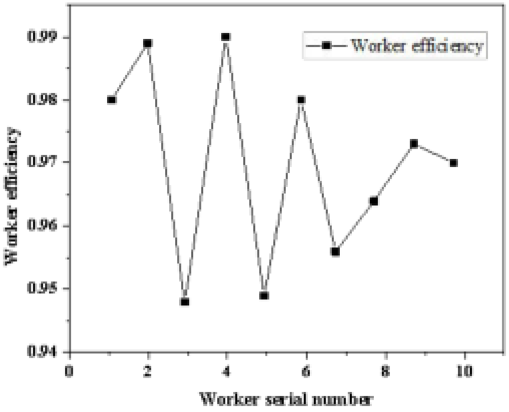


Fig. 4.4: Worker Efficiency

algorithms.

Due to the improved efficiency of workers who choose flexible time, different numbers of flexible workers can also affect the total efficiency of workers under the same unit cost, number of workers, number of tasks, and confidence conditions. Record the changes in the total efficiency of workers when the number of flexible workers is 3, 4, 5, 6, and 7, respectively, while keeping the number of workers, number of tasks, and confidence level unchanged. As shown in Figure 4.3, as the number of flexible workers increases, the total efficiency of workers in different algorithms also increases. In the whale optimization algorithm, the total efficiency of workers approaches the highest value, and thereafter, the growth of the total efficiency of workers slows down. In addition, as the number of flexible workers increases, the worker efficiency of task allocation algorithms based on whale optimization algorithms has always been higher than other algorithms.

In order to discuss the importance of introducing worker flexible time, Figure 4.4 shows the efficiency of each worker in the optimal task allocation results of the Whale Optimization Algorithm. It can be seen that half of the workers have reached the highest efficiency of 1; Figure 4.5 shows the comparison between worker

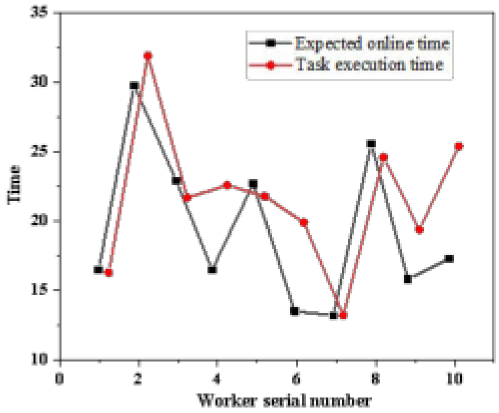


Fig. 4.5: Comparison of Worker Task Execution Time and Estimated Online Time

task execution time and expected online time. It can be seen that workers 2, 4, 6, 9, and 10 chose to extend their time to perform additional tasks, thus achieving the highest efficiency. Therefore, considering worker flexible time can improve worker efficiency.

The efficiency of the workers is close to 1, and the expected online time of the workers is less than 30, and the task execution time of the workers is less than 35.

5. Conclusion. The author proposes a title whale optimization algorithm for efficient task allocation in the Internet of Things, and the task allocation problem is a focus of research related to swarm intelligence perception. In response to the task allocation problem, the author considers the worker's online time as elastic online time and adopts the fuzzy chance constrained programming method for modeling. Due to the fact that the task allocation problem is a combinatorial optimization problem, there is no time efficient optimal solution. Therefore, a two-stage task allocation algorithm was designed based on the whale optimization algorithm for solving. The simulation results show that the task allocation algorithm designed by the author has higher worker efficiency compared to other algorithms.

REFERENCES

[1] Pham, T. M., & Nguyen, T. M. (2024). Function traffic-aware vnf migration for service restoration in an nfv-enabled iot system. ICT Express, 10(2), 374-379.

[2] Abdullaev, I. S., Prodanova, N., Bhaskar, K., Lydia, E., Kadry, S., & Kim, J. (2023). Task offloading and resource allocation in iot based mobile edge computing using deep learning, 76(8), 1463-1477.

[3] Dongre, A. S., More, S. D., Wilson, V., & Singh, R. J. (2024). Medical doctor's perception of artificial intelligence during the covid-19 era: a mixed methods study. Journal of Family Medicine and Primary Care, 13(5), 1931-1936.

[4] Alshanberi, A. M., Mousa, A. H., Hashim, S. A., Almutairi, R. S., Alrehali, S., & Hamisu, A. M., et al. (2024). Knowledge and perception of artificial intelligence among faculty members and students at batterjee medical college. Journal of Pharmacy and Bioallied Sciences, 16(2),1815-1820.

[5] Zhang Yuan, & Zhan Xini. (2022). Research on emergency early warning service mode of emergency crisis events based on ”process-type” intelligence collaboration. Library and Information Service, 66(2), 127-135.

[6] Guo, Y., Yu, T., Wu, J., Wang, Y., Wan, S., & Zheng, J., et al. (2022). Artificial intelligence for metaverse: a framework. CAAI Artificial Intelligence Research, 1(1), 54-67.

[7] Li Guoshan, Y. C. (2022). Is artificial intelligencean aspect-blind? based on wittgenstein's philosophy of perception. Journal of Northeastern University (Social Science), 24(2), 1-7.

[8] Wang, J., Rui, L., Yang, Y., Gao, Z., & Qiu, X. (2023). An incentive mechanism model for crowdsensing with distributed storage in smart cities, 76(8), 2355-2384.

[9] Kun-weiYANG, BoYANG, & Yan-weiZHOU. (2022). A sequential aggregate signature authentication scheme based on blockchain for crowdsensing system. Acta Electronica Sinica, 50(02), 358-365.

[10] Mahmoud, M. M., Atia, B. S., Esmail, Y. M., Ardjoun, S. A. E. M., Anwer, N., & Omar, A., et al. (2023). Application of whale optimization algorithm based fopi controllers for statcom and upqc to mitigate harmonics and voltage instability in modern distribution power grids. Axioms, 29(3), 34-42.

[11] P., C. H., Sujatha, P., & Subbaramaiah, K. (2023). Optimal dg unit placement in distribution networks by multi-objective whale optimization algorithm & its techno-economic analysis. Electric Power Systems Research, 11(03), 261-271.

[12] Wen, C., & Ma, H. (2024). An indicator-based evolutionary algorithm with adaptive archive update cycle for multi-objective multi-robot task allocation. Neurocomputing, 24(2), 259-274.

[13] Alsawadi, M. S., El-Kenawy, E. S. M., & Rio, M. (2023). Advanced guided whale optimization algorithm for feature selection in blazepose action recognition, 37(9), 2767-2782.

[14] Chen, Y., Li, Y., Sun, B., Yang, C., & Zhu, H. (2022). Multi-objective chance-constrained blending optimization of zinc smelter under stochastic uncertainty. Journal of Industrial and Management Optimization, 18(6), 4491-4510.

[15] Gupta, S., Chaudhary, S., Chatterjee, P., & Yazdani, M. (2022). An efficient stochastic programming approach for solving integrated multi-objective transportation and inventory management problem using goodness of fit. Kybernetes, 51(2), 768-803.

[16] Foroutan, R. A., Rezaeian, J., & Shafipour, M. (2023). Bi-objective unrelated parallel machines scheduling problem with worker allocation and sequence dependent setup times considering machine eligibility and precedence constraints. Journal of Industrial and Management Optimization, 19(1), 402-436.

[17] Jalaee, M. S., Ghaseminejad, A., Jalaee, S. A., Zarin, N. A., & Derakhshani, R. (2022). A novel hybrid artificial intelligence approach to the future of global coal consumption using whale optimization algorithm and adaptive neuro-fuzzy inference system. Energies, 15(7), 2578.

[18] Guo, Q., Gao, L., Chu, X., & Sun, H. (2022). Parameter identification for static var compensator model using sensitivity analysis and improved whale optimization algorithm. CSEE Journal of Power and Energy Systems, 8(2), 535-547.

[19] Bin, H. U., Zhu, Y., & Zhou, Y. (2022). Simulated annealing whale radar resource scheduling algorithm based on cauchy mutation. Journal of Northwestern Polytechnical University, 40(4), 796-803.

[20] Dharmalingam, G., Arun, P. M., Panghal, D., Salunkhe, S., Siva, k. M., & Rathinasuriyan, C. (2023). Optimization of awjm process parameters on 3d-printed onyx-glass fiber hybrid composite, 30(2), 84-98.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 11, 2024
Accepted: Jul 18, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1422–1428, DOI 10.12694/scpe.v26i3.4158

INNOVATION OF PRECISION MEDICAL SERVICE MODEL DRIVEN BY BIG DATA

FUJUN WAN[*] XINGYAO ZHOU[†] CHONGBAO REN[‡] AND YUCHEN ZHANG[§]

Abstract. This paper proposes a precision medical service system driven by big data. The PCA-GRA-BK algorithm, which combines principal component analysis (PCA), grey association analysis (GRA) and Bayesian classifier (BK), is adopted. The algorithm extracts critical information from massive medical data, identifies patient characteristics, predicts disease risk, and provides personalized treatment plans. First, the system uses PCA technology to reduce the dimensionality of the original medical data and extract the most representative principal components to reduce data redundancy and retain critical information. Then GRA method was used to analyze the correlation between different medical indicators to determine the main factors affecting health status. Finally, the BK algorithm updates the probability model based on prior knowledge and current data to predict patients' disease risk accurately. A simulation modeling environment is constructed and the PCA-GRA-BK algorithm is tested in this environment to verify the effectiveness of the system. The experimental results show that the algorithm has excellent performance in the accuracy of disease prediction and personalized treatment recommendation. Compared with traditional medical decision support systems, this system has shown significant advantages in extensive data processing capabilities and precision medical services.

Key words: Big data-driven; Precision medicine; PCA-GRA-BK algorithm; Simulation modeling; Personalized treatment.

1. Introduction. In the wave of the digital age, the healthcare field is undergoing an unprecedented transformation. The rise of big data technology has supported the realization of precision medical services. The core concept of precision medicine is individual differences, which emphasize the development of personalized prevention and treatment strategies based on a patient's genetic background, lifestyle and environmental factors. The practice of this concept is inseparable from efficient data processing and analysis technology [1].

In recent years, principal component analysis (PCA), a standard data dimensionality reduction method, has been widely used in the pre-processing stage of medical big data. PCA can transform multiple variables into a few comprehensive variables, thus simplifying the data structure and improving the efficiency of subsequent analysis [2]. Grey correlation analysis (GRA), on the other hand, shows unique advantages in dealing with small samples and uncertainties. By calculating the correlation degree among various factors, GRA reveals the key factors that significantly impact the target [3]. Bayes classifier (BK) is a classification method based on probability statistics. The Bayes classifier (BK) can constantly update the model according to existing data to improve prediction accuracy [3]. The organic combination of these three algorithms to form the PCA-GRA-BK algorithm is expected to provide a comprehensive and efficient set of analytical tools for precision medicine.

Domestic and foreign scholars have made some progress in researching precision medical service systems. Literature [4] proposes a disease prediction model based on deep learning, which can extract features from electronic medical records to achieve early diagnosis of chronic diseases. Literature [5] developed a personalized drug recommendation system based on cloud computing, which used patients' historical medication data to recommend the most appropriate drug combinations. However, most of these studies focus on applying a single algorithm or technology, and lack consideration of the comprehensive performance of the entire precision medical service system.

The research content of this paper aims to build an extensive data-driven precision medical service system based on the PCA-GRA-BK algorithm. First, the paper will elaborate on the principles and steps of the PCA-GRA-BK algorithm and the advantages of their application in medical data processing [6]. Secondly, the

[*] China National Institute of Standardization, Beijing 100000, China
[†] China National Institute of Standardization, Beijing 100000, China (Corresponding author, 18931028393@163.com)
[‡] China Special Equipment Inspection & Research Institute, Beijing 100000, China
[§] China National Institute of Standardization, Beijing 100000, China

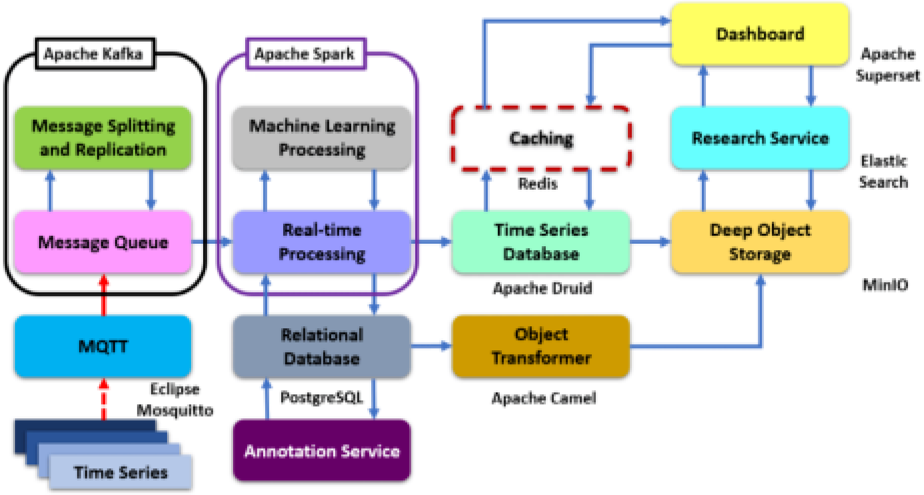


Fig. 2.1: Architecture diagram of distributed cache system for medical big data.

algorithm's performance in the precision medical service system is evaluated through simulation modeling. This paper will simulate different medical scenarios, collect and analyze medical data of virtual patients, and verify the effectiveness of the PCA-GRA-BK algorithm in disease prediction and treatment plan recommendation. Finally, this paper will discuss the potential application of this system in the actual medical environment, as well as the challenges and future development direction.

2. Smart medical data management under big data.

2.1. Research on medical data collection cleaning rules, data warehouse and interface standards. There are many types of medical big data. They exist in relational databases, as well as in unstructured and unstructured parts. Scholars can update and access it in real-time [7]. In addition, due to the characteristics of the medical field itself, it needs to update the status of the data in real time as time goes by. Then, the status of the data ontology is judged in real-time so that all kinds of data can be saved in a unified form. This provides efficient data analysis for both senior and back-office staff. This paper will build an ETL model based on Sqoop. Unified structured data migration between the relational database and the Hadoop platform [8]. Secondly, it uses semi-structured data and unstructured data transfer functions on Hadoop.

A general data interface standard for external applications is proposed. It includes interface format, language, load balancing design, etc. This project plans to develop an intelligent medical cloud computing platform based on Hadoop [9]. In addition, the project adds medical information processing components to improve the operating efficiency of the cloud computing platform. In this way, the real-time acquisition of user physiological parameters, reasonable allocation of resources and directional analysis of display results are realized. Regarding data storage, this paper presents a distributed file management system and a cache database structure. Then, the traditional relational database is supported.

2.2. Design of extensive medical data warehousing and management system.. The storage and management model of medical big data should also meet data warehouse requirements [10]. Then, build a topic-oriented, integrated, changeable and decision-making data warehouse. A distributed Redis architecture based on Zoo Keeper is proposed (Figure 2.1). FIG. 2.1 shows the medical image data processing method. The design is carried out with hierarchical thinking. This architecture divides the overall architecture into two levels. The data layer mainly deals with the specific Redis database and completes the packaging processing of medical services and medical data in the service layer. ZooKeeper is a highly stable performance that ensures efficient cluster load balancing [11]. Zoo Keeper configures multiple backup nodes for Redis hosts. Multiple backups are performed to the Redis host via the Redis backup device to ensure the availability of the Redis cluster. Obtain Redis host slice information from ZooKeeper. The routing method is constructed to solve the problem in the Redis cluster.

3. Medical service data mining analysis and decision-making information service system. The prediction model, association model and service model are studied based on the data model of distributed cache. Integrate it with the needs of intelligent health services to extract practical information from massive data. The scale of data accepted by the system during operation and maintenance is tons, with the rapid data growth [12]. The method studied in this subject can provide an early warning model for developing future diseases and provide a scientific basis for government departments and medical institutions. This project will be based on significant data architecture and medical subject data. They use cutting-edge technologies such as core performance index analysis, cluster gap analysis, data multidimensional analysis, data report analysis, data instrument analysis, etc., for extensive health data analysis and decision support. This project presents a technique for fast storage, indexing and querying massive data. With the continuous expansion of data scale, efficient data storage, indexing and query have become the core problems of data warehousing, and the solution of these problems depends on good data organization and optimization algorithms. A suitable query method is critical to a database. This paper uses collaborative filtering technology to analyze and manage the medical big data stored in the data warehouse. A hospital personalized service platform is constructed using the HL7 communication protocol [13]. The HL7 recommendation message connects the platform with other related software platforms in the hospital.

4. Smart health and medical extensive data display system. In innovative medicine, whether it is patients, doctors, or managers, they want to be able to present the valuable information hidden in big data. In this paper, function mining in intelligent health systems is studied. (1) Enable the hospital to promptly grasp the current medical development trend and adjust the indicators promptly. The medical big data cloud platform enables all hospitals to share medical resources. At the same time, information exchange and collaborative sharing can be conducted promptly [14]. It enables users to obtain service perspectives at multiple levels to achieve the diversity of service models. In addition, the data available on the platform can be used to integrate parts of the medical business. Establish a new service model to save operating costs. (2) Analyze various reports, charts and analysis results. In this way, decisions are made and implemented according to the needs of decision-makers and government staff. (3) Patients conduct a comprehensive analysis and prediction of their case data. Patients can choose the appropriate doctor to consult and get guidance, treatment and reference on the platform. (4) Physicians can also evaluate patients based on patient and platform information. In this way, a personalized treatment plan is developed for the patient.

The ultimate goal of intelligent health management supported by big data is to achieve good interaction on the platform. This paper uses the Zoo Keeper distributed cache framework to build a data representation system for an intelligent health system. Reasonable query for user needs. This project builds a message subscription mechanism based on distributed distribution to meet the diverse processing needs of medical big data [15]. For the different data sources of medical information available, A medical Data processing strategy combining offline batch processing and online real-time computing is proposed (4.1 Cited in How can Big Data Analytics Support People-Centred and Integrated Health Services: A Scoping Review). This project intends to adopt data hierarchical and shunt methods to prolong the data calculation process to minimize the time delay of massive and complex medical big data processing. Then, it is divided into three steps: log parsing, product distribution and new operation. Flink technology and SparkStreaming technology are used in data collection. Flink is a kind of offline parallel stream data processing technology with high throughput and low latency, which is well adapted to the initial log analysis characteristics of medical big data. SparkStreaming is a micro-batch processing method. It can divide the incoming real-time data into several small batches to ensure a stable response during the newly added operations. The added computing delay is minimized to make full use of the efficiency of the computing engine.

5. Gray correlation analysis. Each index's correlation degree is obtained using the grey correlation degree method. In this way, the comprehensive level of each index is constructed. The process is as follows.

1). In studying the grey correlation degree, people must first find the reference sequence reflecting the system's characteristics [16]. Secondly, it is necessary to find out the contrast sequence essential to the whole

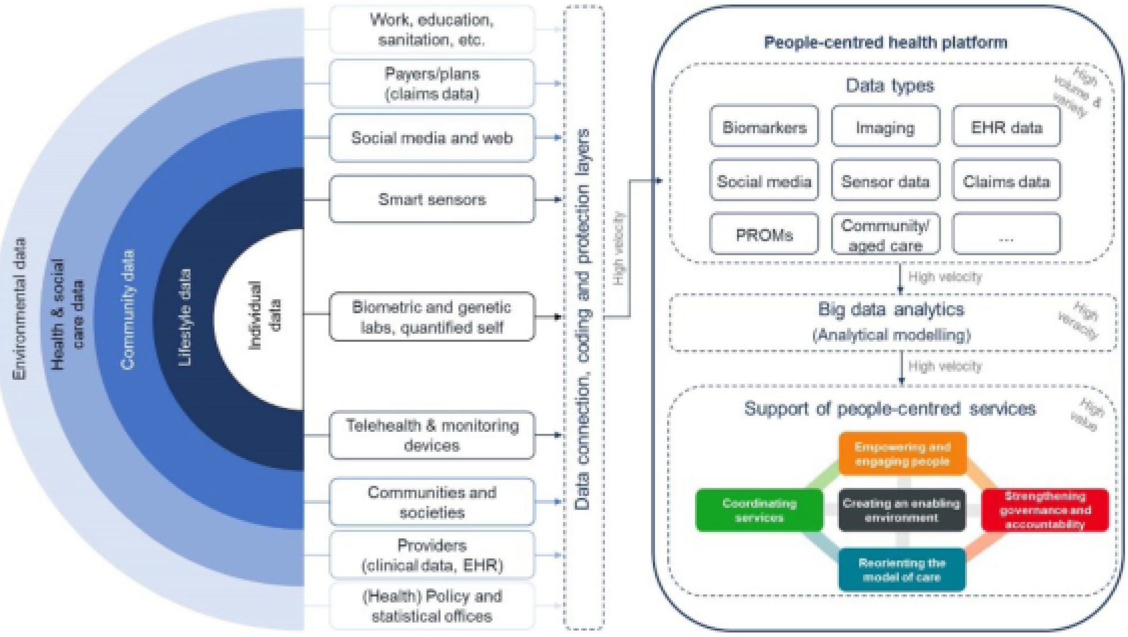

Environmental data
Health & social care data
Community data
Lifestyle data
Individual data
Work, education, sanitation, etc.
Payers/plans (claims data)
Social media and web
Smart sensors
Biometric and genetic labs, quantified self
Telehealth & monitoring devices
Communities and societies
Providers (clinical data, EHR)
(Health) Policy and statistical offices
Data connection, coding and protection layers
High velocity
People-centred health platform
Data types
High volume & variety
Biomarkers
Imaging
EHR data
Social media
Sensor data
Claims data
PROMs
Community/ aged care
...
High velocity
Big data analytics
(Analytical modelling)
High veracity
High velocity
Support of people-centred services
High value
Empowering and engaging people
Coordinating services
Creating an enabling environment
Strengthening governance and accountability
Reorienting the model of care

Fig. 4.1: Medical big data processing strategy.

system. A set of data that can reflect the characteristics of a system is called a reference sequence, which is used to measure and compare the degree of correlation between the sequences. Sensitivity is usually treated as A reference sequence $F = F(t) \mid 1, 2, \cdots, n.F(t)$ represents the number that corresponds to the reference sequence. A series of factors is called a relative sequence. It can be expressed in terms of $G_i = G_i(t) \mid t = 1, 2, \cdots, n, i = 1, 2, \cdots, m.g_i(t)$ represents the t th value in the i th comparison series, and n represents the number of QI features.

2). Because the measurement units are not necessarily the same, it is necessary to average them first when conducting gray correlation analysis. Take the average value of each value in the sequence so that the processed data value is close to the order of 1 .

$$g_i'(t) = \frac{g_i(t)}{g_i} \tag{5.1}$$

$\bar{g}_i$ is the average of series i. $g_i(t)$ means that the t data in the i order is averaged. $g_i(t)$ is the t data after the average processing of the i data.

3). the quantity difference between the various data is reduced in the standardization process. This makes the calculation of the grey correlation coefficient more convenient. Its expression is as follows:

$$\lambda_i(t) = \frac{\min_i \min_t |y(t) - g_i(t)| + \delta \cdot \max_i \max_t |y(t) - g_i(t)|}{|y(t) - g_i(t)| + \delta \min_i \min_t |y(t) - g_i(t)|} \tag{5.2}$$

$|y(t) - g_i(t)|$ is the distance between the reference sequence and the corresponding t data in the i contrast sequence, where max is the most significant distance and min is the shortest distance. δ is also known as the distinguishing factor, and the value range of δ is usually (o,1).

4). The correlation factor reflects the degree of correlation between each comparison sequence and the baseline sequence [17]. When measuring correlation degree, the correlation coefficient should be used to calculate the average value of different time points. Relevance r_i is calculated as follows:

$$SSE = \sum_{i=1}^{t} \sum_{q \in Z_i} |q - n_i|^2 \tag{5.3}$$

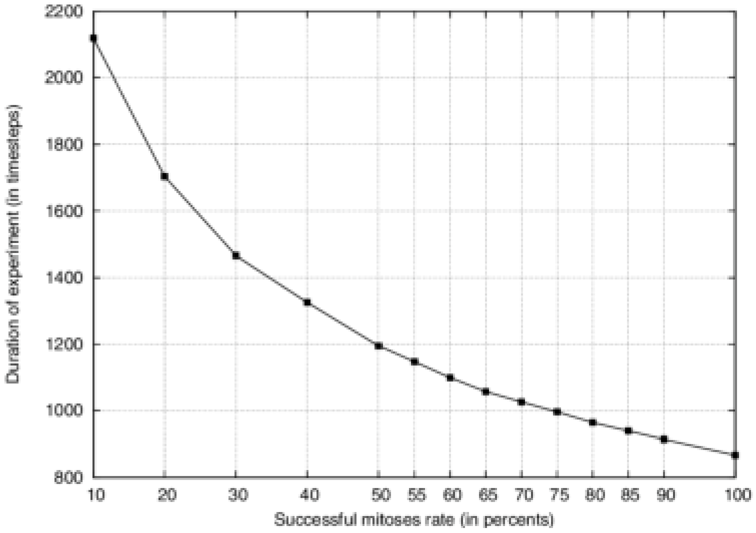


Fig. 7.1: Elbow chart.

When the desired correlation degree is closer to the same time, the higher the correlation degree between the sequences, the higher the correlation degree between the identifier property and the sensitive data.

6. Grey Correlation analysis of medical big data (PCA-GRA-BK algorithm). This project studies the PCA-GRA-BK method based on privacy and validity. The GRA method was used to evaluate each quality index's correlation degree to determine the comprehensive level applicable to each index [18]. The quality information with maximum value must be selected for processing in data collection, thus extending the universality of quality information. First, it must be divided into categories to prevent the K-type anonymous system from being too general. An adaptive method is used to select the number of categories and K-anonymize them. This improves the efficiency of classification and reduces the packet loss rate. Records satisfying K anonymizers are selected from set T and added to K hidden tables [19]. The value of the optimal class n represents the number of clusters of a class. The sample set K is divided into n samples, and the maximum quasi-recognition item of the sample set is found.

7. Experimental analysis. This project intends to use the elbow method to determine the optimal number of classes. By calculating the sum of error squares (SSE) between classes, class families and SSE values are taken as coordinates by points [20]. The number of optimal clusters is determined based on the value of the inflection point closest to the shape of the elbow.

$$SSE = \sum_{i=1}^{t} \sum_{q \in Z_i} |q - n_i|^2 \tag{7.1}$$

Z_i is the ith cluster, q is the sampling point in Z_i, and n_i is the average of all samples on Z_i. According to point A's coordinates, the elbow stroke method is used to compare the test results and get a reasonable number of clusters. It can be seen from Figure 7.1 that when the number of groups is 6 , the judgment criterion of the elbow method is satisfied, so the optimal number of classes in this data set is set to 6. Then this paper carries out cluster analysis based on the optimal class size. The performance of Datafly, PCA-GRA Datafly, PCA-GRA-KK and PCA-GRA-KK algorithms is compared and analyzed. The amount of information is an important index to evaluate the algorithm's performance, and the loss degree of this index is small, indicating that a lot of original data is lost in the system [21]. It has a high use efficiency. The calculation method of

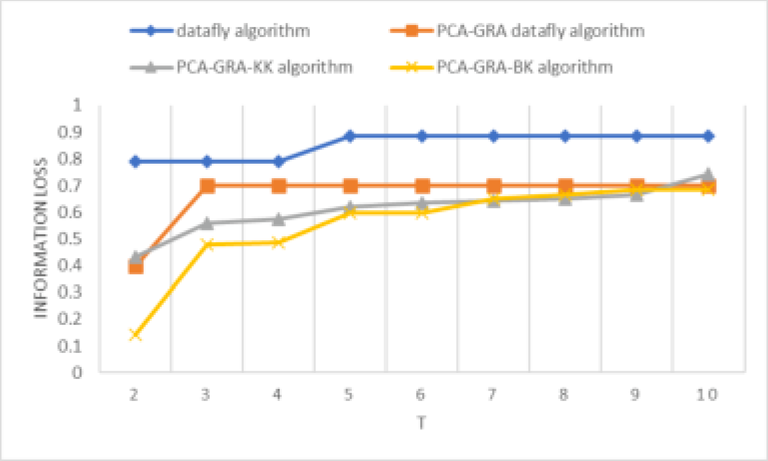


Fig. 7.2: Graph of changes in information loss rate.

information loss is given:

$$IL(RT) = \frac{\sum_{i=1}^{n_q} \sum_{j=1}^{n} \frac{L}{|DHG_A|} + \sum_{t=1}^{n} |k[Q]|}{|K| \times n_q} \tag{7.2}$$

K is the initial data table, RT is the anonymized data table, n is the number of tuples included in the data table, n_q is the number of quasi-identifier features, L is the number of times that the class j identifier of the i record is promoted in the generalization tree, $|DHG_A|$ is the height of the generalization tree of feature A, $|k[Q]|$ is the value of the class identifier Q in the record. The PCA-GRA-KK method is combined with the PCA-GRA-BK method to cluster groups with high similarity. Then, local generalization and K-anonymization are performed to reduce the data loss caused by the overall generalization. The PCA-GRA-BK method is used to optimize the clustering, thus reducing the packet loss rate of the grouping. The critical problem in anonymizing medical data is preserving the original data's information as much as possible. The results showed the best PCAGRA-BK method (Figure 7.2).

8. Conclusion. This study successfully constructed an extensive data-driven precision medical service system based on the PCA-GRA-BK attracted. By integrating principal component analysis (PCA), grey association analysis (GRA) and Bayesian classifier (BK), the system realized rapid processing and accurate analysis of large-scale medical data. The system has shown good performance and accuracy in disease risk prediction and personalized treatment plan recommendation through simulation modeling. The experimental results show that the PCA-GRA-BK algorithm can effectively extract critical information from complex and changeable medical data, identify the core factors affecting health, and constantly improve the accuracy of disease prediction through the dynamic updating characteristics of the Bayes classifier. In addition, the personalized treatment recommendation function of the system can provide more suitable treatment plans based on considering the specific situation of patients, thus improving the pertinence and effectiveness of medical services. However, although this research has achieved positive results, it still needs to face challenges in practical applications such as data privacy protection, algorithm transparency, and system stability. Future research efforts will further optimize the algorithm, strengthen data security and privacy protection measures, and explore more clinical application scenarios to ensure the system's sustainable development and broad application.

REFERENCES

[1] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of big data, 6(1), 1-25.

[2] Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics, 15(2), 94-101.

[3] Singh, R. P., Javaid, M., Haleem, A., & Suman, R. (2020). Internet of things (IoT) applications to fight against COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 521-524.

[4] Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339.

[5] Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1328-1347.

[6] Sreenu, G. S. D. M. A., & Durai, S. (2019). Intelligent video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data, 6(1), 1-27.

[7] Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural computing and applications, 32(24), 18069-18083.

[8] Liu, H., Ong, Y. S., Shen, X., & Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. IEEE transactions on neural networks and learning systems, 31(11), 4405-4423.

[9] Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. British journal of management, 30(2), 272-298.

[10] Kumar, S., Tiwari, P., & Zymbler, M. (2019). Internet of Things is a revolutionary approach for future technology enhancement: a review. Journal of Big data, 6(1), 1-21.

[11] Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big data, 6(1), 1-18.

[12] Tian, S., Yang, W., Le Grange, J. M., Wang, P., Huang, W., & Ye, Z. (2019). Smart healthcare: making medical care more intelligent. Global Health Journal, 3(3), 62-65.

[13] Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big data, 6(1), 1-16.

[14] Javaid, M., & Khan, I. H. (2021). Internet of Things (IoT) enabled healthcare helps to take the challenges of COVID-19 Pandemic. Journal of oral biology and craniofacial research, 11(2), 209-214.

[15] Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature medicine, 25(1), 37-43.

[16] Qadri, Y. A., Nauman, A., Zikria, Y. B., Vasilakos, A. V., & Kim, S. W. (2020). The future of healthcare internet of things: a survey of emerging technologies. IEEE Communications Surveys & Tutorials, 22(2), 1121-1167.

[17] Niebel, T., Rasel, F., & Viete, S. (2019). BIG data–BIG gains? Understanding the link between big data analytics and innovation. Economics of Innovation and New Technology, 28(3), 296-316.

[18] Shen, M., Deng, Y., Zhu, L., Du, X., & Guizani, N. (2019). Privacy-preserving image retrieval for medical IoT systems: A blockchain-based approach. Ieee Network, 33(5), 27-33.

[19] Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of travel research, 58(2), 175-191.

[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] Santos, M. K., Ferreira, J. R., Wada, D. T., Tenório, A. P. M., Nogueira-Barbosa, M. H., & Marques, P. M. D. A. (2019). Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. Radiologia brasileira, 52(06), 387-396.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 21, 2024
Accepted: Jul 25, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1429–1439, DOI 10.12694/scpe.v26i3.4167

BUILDING ENERGY SYSTEMS USING DIGITAL TWINS AND GENETIC ALGORITHMS

JIFENG HAN[*] AND LIQIN AI[†]

Abstract. In order to solve the problems of energy consumption behavior and production process generating heat waste and carbon emissions, the author proposes to use digital twins and genetic algorithms to study building energy systems. The author employed Matlab/Simulink to develop an optimization framework for isolated multi-energy complementary building energy systems. The optimization objective was to minimize the annual cost of the system, and based on digital twins and genetic algorithms, the model was optimized and simulated for analysis. The experimental results show that compared to not considering flexible loads, when flexible electrical loads, flexible thermal loads, and flexible electrical/thermal loads participate in regulation, the annual cost of the system is reduced by 5.13%, 33.01%, and 35.4%, respectively. Incorporating flexible electrical loads into regulation shifts energy demand towards periods of high photovoltaic output, thereby reducing the required capacities of energy storage batteries and diesel generators. Compared to scenarios where only flexible thermal loads participate in regulation, simultaneous participation of both flexible electrical and thermal loads results in smoother indoor temperature fluctuations with reduced amplitude. When flexible thermal and electrical loads are simultaneously regulated, the best effect is achieved in reducing the annual value of system costs and annual carbon dioxide emissions.

Key words: Multi energy complementarity, Flexible load, Digital twin, Isolated building energy systems, genetic algorithm

1. Introduction.

As national society progresses, industrialization advances, and living standards rise, the demand for energy continues to grow steadily [1]. The total energy consumption of building energy systems is high and is influenced by weather, indoor personnel behavior, and comfort needs. It mainly consumes a large proportion of energy in cooling, heating, and lighting systems. Therefore, the optimization of building energy system operation and energy-saving strategies have received widespread attention from scholars [2].

The traditional energy consumption method usually uses manual methods to predict resources and optimize scheduling in real-time based on the predicted resources; This method has the disadvantages of many uncertain factors, a small range of scheduling strategy selection, and large prediction errors [3]. For example, due to information asymmetry among users, between users and the power grid, or between users, energy conservation and emission reduction goals cannot be achieved, and the requirements for equipment performance among users are inconsistent or even deviate from the goals; It is difficult to achieve when resources are optimized and scheduling strategies are configured to achieve this goal; The energy consumption behavior and production process generate problems such as heat waste and carbon emissions that are difficult to eliminate [4]. Therefore, in order to further achieve the goal of efficient energy conservation and emission reduction, it is necessary to conduct unified analysis and optimization of various resources, and dynamically analyze and simulate them to obtain scheduling strategies with universal laws and good targeting and energy-saving effects. By using digital twin and genetic algorithm technology, a dynamic energy efficiency model is constructed and applied to the overall planning of the comprehensive energy platform to achieve goals such as improving energy utilization efficiency and reducing energy consumption. Establishing a multi energy complementary building energy system that couples renewable energy and traditional energy can effectively overcome the intermittency and volatility of renewable energy, which is of great significance for ensuring the reliability and stability of energy supply in isolated rural areas and achieving local energy self-sufficiency [5,6].

2. Literature Review.

Due to the intermittent and fluctuating characteristics of renewable energy, it is difficult to match the supply and demand of multi energy complementary building energy systems. The integration of energy storage devices is an important technology for achieving supply-demand balance in re-

[*] Nanchang Institute of Science & Technology, Nanchang, 330108, China.
[†] Nanchang Institute of Science & Technology, Nanchang, 330108, China.(Corresponding author, aili118899@163.com)

newable energy systems at present, but relying solely on energy storage devices for regulation will greatly increase system costs [7-8]. Flexible load regulation refers to a type of method that optimizes the load curve by actively changing the load operating time or load size, and has become a hot research topic for domestic and foreign industry scholars due to its economic and efficient characteristics [9]. Libralato, M. et al. created a digital twin to analyze the energy consumption of building HVAC systems. They detailed the programming and data analysis framework of the supervisory system. The digital twin was then employed to compare two control strategies for summer thermostat regulation, aiming to enhance the energy efficiency of building HVAC systems and leverage the thermal storage properties of building envelope structures to modify and reduce peak power demand [10]. Ohmura, T. et al. presented a use case examining optimal scheduling and energy-saving parameters. The study revealed that certain parameter configurations could reduce work waiting time by up to 70% and decrease energy consumption by 1.2% during peak system activity. Consequently, this digital twin demonstrated the feasibility for system administrators to accurately adjust various parameters without disrupting system operations [11]. Hou, Y. et al. introduced a combined simulation framework for energy auditing and pixel-level simulation of building envelope structures that integrates with Digital Twins (DT). This framework initially examines the input and output interactions between the Building Physics Twin (PT) and DT for energy auditing, highlighting the current technical challenges in transferring data from PT to DT for building energy simulation. Additionally, it evaluates the requirements for building parameters in simulations and identifies the existing research gap in model updates between Building Information Modeling (BIM) and Building Energy Modeling (BEM). Furthermore, the framework presents joint simulation methods for building energy simulation, detailing how to exchange data within the joint simulation and interpret the results to develop renovation plans [12].

The above research comprehensively analyzed the impact of digital twins and genetic algorithms on the capacity configuration of energy system equipment, but the research objects are mostly limited to multi input single output microgrid systems. The author aims to establish a matching mechanism between multi type heterogeneous energy combination supply and flexible demand, establish an isolated multi energy complementary building energy system design optimization model for solving multi input and multi output, and analyze and summarize the effect of flexible electricity/heat load on the overall performance of the system through specific cases. The research results can lay a theoretical foundation for the planning and design of distributed multi energy complementary building energy systems in villages.

3. Method.

3.1. Principle of Thermal Power Supply. The isolated multi energy complementary building energy system consists of photovoltaic modules, solar collectors, air source heat pumps, diesel generators, energy storage batteries, and heat storage water tanks (see Figure 1). The system can be divided into two major components: Power supply system and heating system. In the power supply system, photovoltaic modules serve as the main power supply equipment, and diesel generators serve as auxiliary power supply equipment to supply power to air source heat pumps, electric heaters, and buildings [13,14]; In the heating system, solar collectors serve as the main heating equipment, while air source heat pumps and electric heaters serve as auxiliary heating equipment to provide heat to buildings. The energy storage equipment of the system consists of a heat storage water tank and energy storage batteries.

3.2. Heating control strategy. The operation of a solar collector is influenced by the temperature difference between its inlet and outlet, as well as the maximum temperature of the heat storage tank. Similarly, the start and stop of an air source heat pump and electric heater are affected by the upper temperature limit of the heat storage tank. The heating control strategy is illustrated in Figure 3.2. In this figure, the upper temperature limit of the heat storage tank is set at $50°C$. The temperature differences for starting and stopping the solar collector are $5°C$ and $2°C$, respectively. For the air source heat pump, the start and stop temperatures are $40°C$ and $45°C$, while the start temperature for the electric heater is $40°C$.

3.3. Power supply control strategy. The power supply strategy for isolated multi energy complementary building energy systems is shown in Figure 3.3. When the power generation of photovoltaic modules exceeds the sum of air source heat pumps, electric heating, and building electricity consumption (total electricity consumption), only photovoltaic modules are used for power supply; When the power generation of

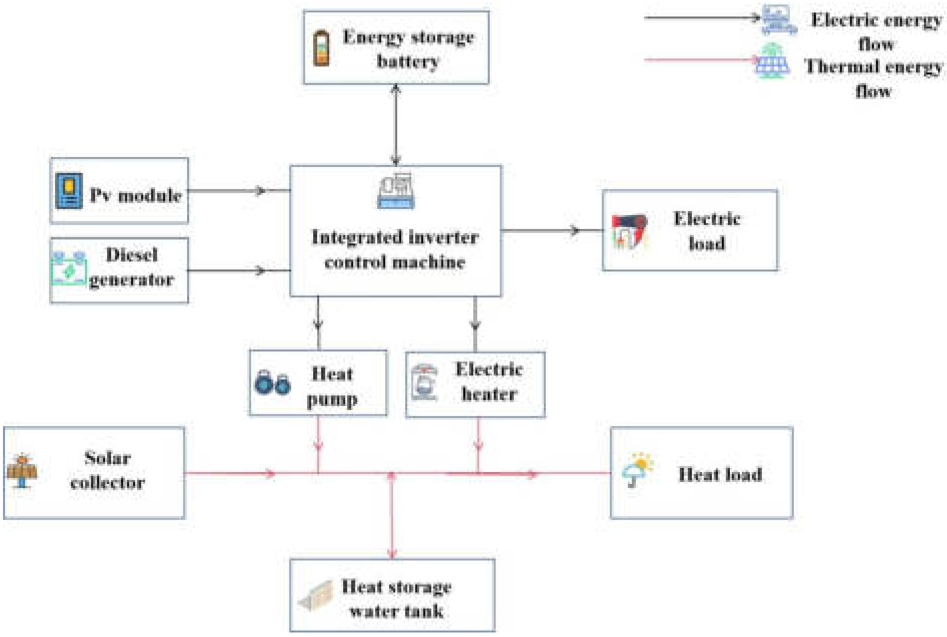


Fig. 3.1: Isolated Multi energy Complementary Building Energy System

photovoltaic modules is less than the total electricity consumption and has not yet reached the lower limit of energy storage battery discharge, priority should be given to the auxiliary photovoltaic modules powered by energy storage batteries. After reaching the lower limit of energy storage battery capacity, diesel generators should be started for power supply [15].

3.4. Establishment of mathematical models for equipment.

3.4.1. Solar collectors. The author utilizes flat plate solar collectors as the primary heating equipment. The heat collection capacity of a solar collector depends on the solar irradiance and the effective area of the collector [16]. In a stable state, ignoring the heat absorbed by the heat absorbing plate itself, the formula for calculating the solar heat collection is:

$$Q_{cu} = A_{co} \cdot [F_R \cdot (\tau\alpha)_e \cdot G - F_R \cdot U_L \cdot (T_{ci} - T_{en})] \tag{3.1}$$

In the above equation: Aco is the effective area of the solar collector, m^2; T_{ci} is the inlet temperature of the solar collector, $^\circ C$; Ten ambient temperature, $^\circ C$.

The heat collection of a solar collector can also be expressed by the inlet and outlet temperature of the solar collector, and the calculation formula is:

$$Q_{cu} = 3600 \cdot c \cdot m_{co} \cdot (T_{co} - T_{ci}) \tag{3.2}$$

In the above formula: T_{co} solar collector outlet temperature, $^\circ C$; T_{ci} is the inlet temperature of the solar collector, $^\circ C$.

3.4.2. Heat storage water tank. The author employs the 'node method' to model the thermal stratification phenomenon in the water tank. The control functions for the solar collector side and the load side are as follows:

$$F_i^c = \begin{cases} 1, i = 1, T_{co} > T_i \\ 1, T_{i-1} \geqslant T_{co} > T_i \\ 0, i = 0 || i = N+1 \\ 0, other \end{cases} \tag{3.3}$$

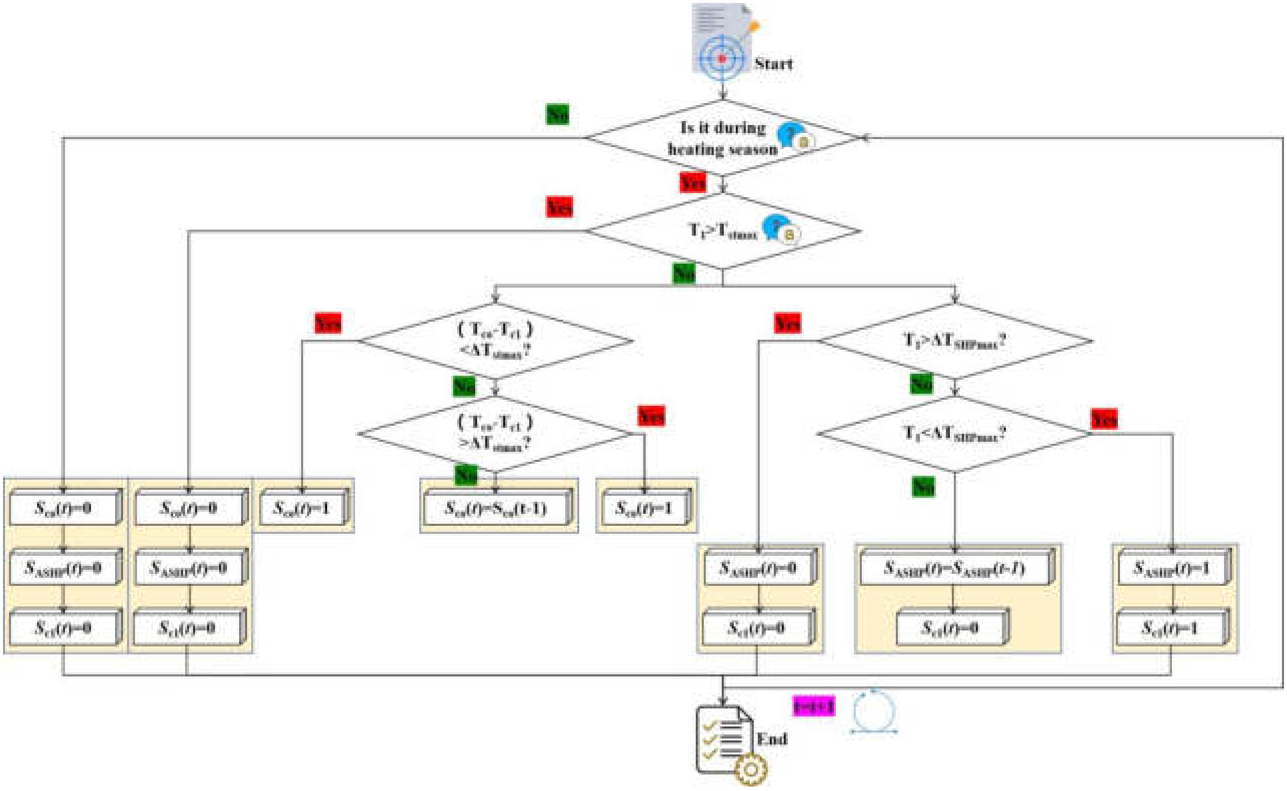


Fig. 3.2: Heating System Control Strategy

$$F_i^L = \begin{cases} 1, i = N, T_L < T_N \\ 1, T_{i-1} \geqslant T_L > T_i \\ 0, i = 0 || i = N+1 \\ 0, other \end{cases} \tag{3.4}$$

In the above equation: T_i represents the average water temperature at node i of the thermal storage tank, $°C$. T_L is the return water temperature on the load side, $°C$; T_N is the average water temperature (lowest layer) of node N in the thermal storage tank, $°C$.

The energy balance relationship of node i is as follows:

$$\begin{gathered} m_i \cdot \frac{dT_i}{dt} = [\frac{UA}{c_p} \cdot (T_{en} - T_i) + F_i^c \cdot m_{co} \cdot (T_{co} - T_i) + \\ F_i^L \cdot m_L \cdot (T_L - T_i) + \begin{cases} \dot{m_i} \cdot (T_{i-1} - T_i), \dot{m_i} > 0 \\ \dot{m_{i+1}}(T_i - T_{i+1}), \dot{m_{i+1}} < 0 \end{cases} \quad] \end{gathered} \tag{3.5}$$

3.4.3. Air source heat pump. The author used a fitted coefficient of performance (COP) curve to calculate the heating capacity of an air source heat pump, taking into account the defrosting correction of the air source heat pump in the study. The calculation formula is:

$$COP = -0.0004 \cdot T_{en}^2 + 0.0903 \cdot T_{en} + 3.0924 \tag{3.6}$$

$$Q_{ASHP} = S_{ASHP} \cdot P_{ASHP} \cdot COP \cdot (1 - k) \tag{3.7}$$

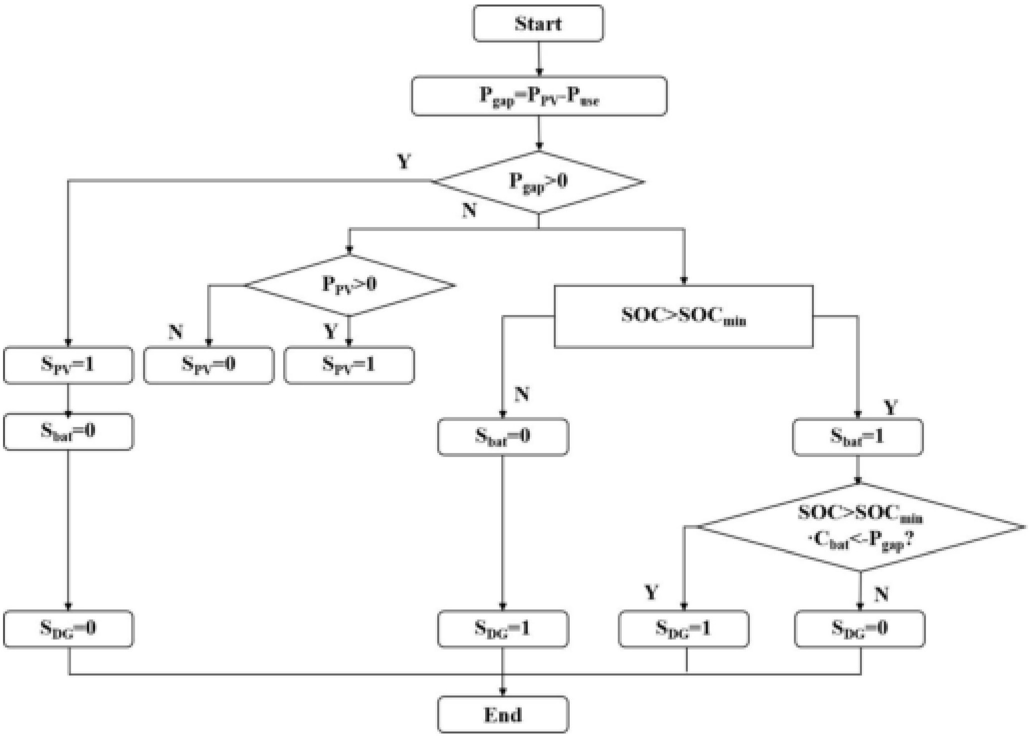


Fig. 3.3: Power Supply System Control Strategy

3.4.4. Photovoltaic modules. The formula for calculating the power generation of photovoltaic modules is:

$$P_{PV} = PMP_{ref} \cdot \frac{G}{1000} \cdot [1 + \gamma_{PV} \cdot (T_{cell} - 25)] \tag{3.8}$$

The above equation: γ_{PV} is the temperature coefficient of the photovoltaic module's power generation efficiency, ranging from 0.4% to 0.6%/$°C$; T_{cell} is the battery temperature, $°C$.

The battery temperature T_{cell} can be calculated by equation 3.9:

$$T_{cell} = T_{en} + \frac{NOCT - 20}{800} \cdot G \tag{3.9}$$

3.4.5. Energy storage batteries. The State of Charge (SOC) calculation during the charging and discharging process of energy storage batteries is as follows:

$$SOC(t) = SOC(t - \Delta t) + \frac{P_{ban} \cdot \eta_{cha} \cdot \Delta t}{C_{bat}} \tag{3.10}$$

$$SOC(t) = SOC(t - \Delta t) + \frac{P_{baT} \cdot \Delta t}{\eta_{dis} \cdot C_{bat}} \tag{3.11}$$

The above equation: SOC(t) t represents the charging and discharging status of the energy storage battery at time t; η_{cha} is the charging efficiency of the energy storage battery, taken as 1; η_{dis} is the discharge efficiency of the energy storage battery, taken as 0.8.

3.4.6. Diesel generator. As an emergency power supply, the fuel consumption of diesel generators depends on their rated power and output power. The approximate mathematical expression is:

$$F_{cons} = \alpha_{DC} \cdot P_{DG} + \beta_{DG} \cdot P_{o-DG} \tag{3.12}$$

The above equation: α_{DC} is the coefficient of diesel generator, taken as 0.081451/kWh; β_{DG} is the coefficient of diesel generator, taken as 0.2461/kWh.

3.5. Electric load model. The author divides the user side electricity load into three categories: 1) Basic electricity load: closely related to the living habits of residents, and cannot change their energy consumption mode and time; 2) Translatable electrical load: The power supply time of the load can be changed, but the load needs to be moved as a whole and cannot be interrupted [17]. 3) Transferable electricity load: The electricity consumption during each time period can be flexibly adjusted, but it must ensure that the total load of the entire cycle remains unchanged after the transfer compared to before the transfer. The specific modeling of various flexible electrical loads is detailed in the following text [18].

3.5.1. Translatable electrical load. Assuming a unit scheduling time of 1 hour, for the translatable electrical load Lmove, the power distribution before participating in scheduling is expressed as:

$$L^{*}_{move} = [0, \cdots, P_{more}(t_s), P_{more}(t_s + 1), \cdots, P_{move}(t_s + t_b), \cdots, 0] \tag{3.13}$$

Assuming the translatable interval is $[t_{move}, t_{move}]$, use the 0-1 variable a to represent the translational state of L_{move} at a certain time period t. When a=1, it indicates that Lmove starts translational from time t; when a=0, it indicates that L_{move} does not translational. The set of starting time periods S_{move} is:

$$S_{move} = [t_{move-}, t_{move} - t_D + 1] \tag{3.14}$$

If $t \in [t_{mox}, t_{moxot} - t_D + 1]$ and $t \neq t_s$, then the power distribution of L^{*}_{max} shifting from time t ϵ to L_{max} at time t is:

$$L_{move} = [0, \cdots, P_{move}(t), P_{move}(t + 1), \cdots, P_{mover}(t + t_D), \cdots, 0] \tag{3.15}$$

3.5.2. Transferable electrical load. Assuming that the transferable interval of the transferable electrical load L_{trm} is $[t_{tan}, t_{trmn}]$, the 0-1 variable b represents the transfer state of L_{trom} at time t, and b(t)=1 represents the transfer of power P_{tan} in L_{trn} at time t. The power constraints after the transfer are as follows:

$$b(t) \cdot P^{tran}_{min} \leqslant P_{tran}(t) \leqslant b(t) \cdot P^{tran}_{max} \tag{3.16}$$

If there are no restrictions on the load transfer period, there may be a phenomenon of load transfer to multiple single periods, which is manifested externally as frequent equipment startup and shutdown. Therefore, it is necessary to limit the minimum duration of load transfer operation:

$$\sum_{\tau=t}^{+T^{tman}_{mimen}-1} b(\tau) \geqslant T^{tran}_{min} \cdot (b(\tau) - b(\tau - 1)) \tag{3.17}$$

By using the above model, the adjusted flexible electrical load can be obtained, and the total electrical load can be calculated as follows:

$$P_{active} = P_{move} + P_{tran} \tag{3.18}$$

$$P_{load} = P_{base} + P_{acctive} \tag{3.19}$$

3.6. Heat load model.

3.6.1. Calculation of building heat load. The heat load calculation model for the room is:

$$Q_{load} = q_V \cdot V \cdot (T_{in} - T_{en}) \tag{3.20}$$

3.6.2. Heat load model considering flexible loads. Assuming that the indoor temperature varies between and the variable di represents the difference between the upper limit of the indoor temperature and the actual indoor temperature, the indoor temperature is:

$$T_{in}(t) = T_{inmax} - d_i(t) \tag{3.21}$$

By substituting formula 3.21 into formula 3.20, a heat load model considering flexible loads can be obtained [19].

3.7. Optimization Model.

3.7.1. Objective function. The author sets the annual cost of isolated multi-energy complementary building energy systems as the objective function for optimization. The annual cost of the system comprises two components: the annualized investment value and the operation and maintenance costs. The mathematical expression is:

$$min(F) = min(C_1 + C_{om}) \tag{3.22}$$

The mathematical expression for the annual value C1 of system investment is:

$$C_1 = \left[P_h \cdot \frac{i_{ATT}}{1-(1+i_AIT)-T_h} + P_e \cdot \frac{i_{ATT}}{1-(1+i_{AIT}-T_c)} + C_{bat} \cdot \frac{i_{AIT}}{1-(1+i_{AIT})-T_{Lac}}\right] \tag{3.23}$$

$$P_h = C_{1\_co} \cdot A_{co} + C_{1ASHP} \cdot P_{ASHP} + C_{1\_st} \cdot V_{st} \tag{3.24}$$

$$P_e = C_{1\_PV} \cdot PMP_{ref} + C_{1\_DG} \cdot P_{DG} + C_{1\_acc} \tag{3.25}$$

$$C_{bat} = C_{1Lhat} \cdot C_{bat} \tag{3.26}$$

The operating cost of the system is the fuel cost generated by diesel power generation, and the mathematical expression for operating and maintenance costs is:

$$C_{om} = C_F \cdot F_{cons} + C_1 \cdot \zeta \tag{3.27}$$

3.7.2. Constraints. The capacity and operation of each device in the system should be within a certain reasonable range, and the corresponding equipment capacity and power range are:

$$0 \leqslant A_{co} \leqslant A_{co,max} \tag{3.28}$$

$$0 \leqslant V_{st} \leqslant V_{st,max} \tag{3.29}$$

$$0 \leqslant PMP_{ref} \leqslant PMP_{ref,max} \tag{3.30}$$

$$0 \leqslant C_{hat} \leqslant C_{hat,max} \tag{3.31}$$

$$0 \leqslant P_{o-DG} \leqslant P_{DG} \tag{3.32}$$

$$0 \leqslant P_{ASHP} \leqslant \frac{Q_{load,max}}{COP} \tag{3.33}$$

$$SOC_{min} \leqslant SOC \leqslant SOC_{max} \tag{3.34}$$

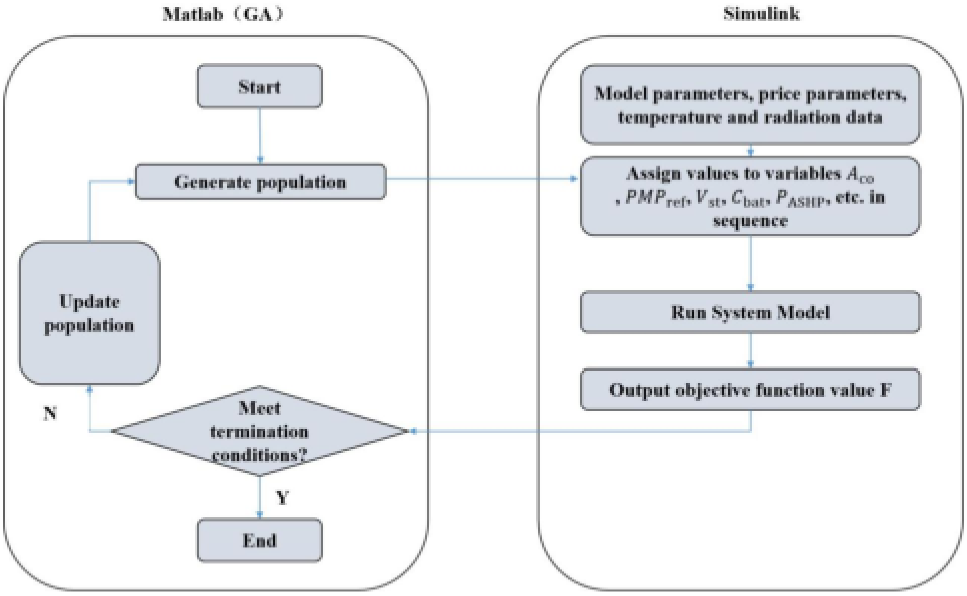

Matlab（GA）
Simulink
Start
Generate population
Update population
N
Meet termination conditions?
Y
End
Model parameters, price parameters, temperature and radiation data
Assign values to variables A_{co}, PMP_{ref}, V_{st}, C_{bat}, P_{ASHP}, etc. in sequence
Run System Model
Output objective function value F

Fig. 3.4: System Optimization Flowchart

Table 3.1: Unit Costs of System Equipment

parameter	numerical value	parameter	numerical value
$C_{I\_co}/(yuan \cdot m^{-2})$	800	$C_{I\_ASHP}/(yuan \cdot KWW^{-1})$	2000
$C_{I\leqslant 1}/(yuan \cdot m^{-3})$	500	$C_{I\_PV}/(yuan \cdot KWW^{-1})$	8000
$C_{.1bat}/(yuan \cdot kWh^{-1})$	800	$C_{I\_DG}/(yuan \cdot KWW^{-1})$	1500

3.7.3. Optimization methods. The author established a system model using Simulink, and based on genetic algorithm, jointly solved it using Simulink and Matlab optimization toolbox. The optimization process is shown in Figure 3.4. At the beginning of the iteration, the decision variables are passed to the Simulink module, and the objective function value is calculated using dynamic simulation of the system and returned to MATLAB for judgment [20]. If the termination condition is satisfied, the iteration stops and the optimal result is achieved; otherwise, the iteration continues until the termination condition is met.

3.8. Input parameter settings. The author studied calculating a time step of 1 hour, setting different population sizes and maximum iterations for different scenarios, with a maximum population size of 1490 and a maximum iteration of 400. The unit costs of each equipment involved in the calculation process are shown in Table 3.1.

3.9. Experimental Analysis. The author takes a rural residential building as the research object, with a heating area of 68m2 and an indoor calculated temperature of 18 $°C$.

The daily electrical load of buildings includes basic electrical load, translatable electrical load 1, translatable electrical load 2, and transferable electrical load. The parameters of each flexible load are shown in Table 3.2. The author's research assumes that the hourly electricity consumption of the building is the same during winter, transition season, and typical summer days.

The author studied using annual cost values as performance evaluation indicators for isolated multi energy complementary building energy systems, with carbon dioxide emissions as auxiliary performance evaluation

Table 3.2: Flexible Load Parameters

Types	Translatable electrical load		
	t_s t_D/h $t_{mone} \sim t_{monet}$	t_s t_D/h $t_{mone} \sim t_{monet}$	t_s t_D/h $t_{mone} \sim t_{monet}$
1	19:00	3	08:00-20:00
2	11:00	2	08:00-20:00
Transferable electrical	T_{min}^{tam}/h	$P_{min} \sim P_{man}/kW$	$t_{tran} \sim t_{tran}$
load	2	$0.15 \sim 0.25$	$08:00-20:00$

Table 4.1: Optimization Results of Multi energy Complementary Building Energy System Design

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4
A_{co}/m^2	16.17	11.72	9.70	9.45
P_{ssH}/kW	2.60	2.07	1.31	2.06
V_s/m^3	9.84	9.87	7.88	8.47
PMP/kW	2.03	2.00	1.88	1.81
C_{ba}/kWh	68.87	71.26	66.60	64.43
P_{Dc}/kW	0.05	0	0.04	0
F/ ten thousand yuan	2.63	2.50	1.72	1.66
E_{Co_2}/kg	120.35	0	92.13	0
Cost savings rate/%	-	5.13	33.00	35.40

indicators. The formula for calculating the annual carbon dioxide emissions of diesel generators is:

$$E_{CO_2} = \sum_{t=1}^{8760} F_{cons}(t) \cdot EF \tag{3.35}$$

4. Results and Discussion. To analyze the impact of flexible loads on the performance of multi-energy complementary building energy systems, the author establishes four scenarios for comparative analysis: Scenario 1, which does not consider flexible load regulation; Scenario 2, which considers flexible electrical load regulation; Scenario 3, which considers flexible heat load regulation; and Scenario 4, which considers both flexible electrical and thermal load regulation. The optimization results for the multi-energy complementary building energy system under these four scenarios are presented in Table 4.1.

From Table 4.1, it can be seen that in Scenario 1, the solar collector, air source heat pump, and diesel generator have the highest capacity. In scenario 2, the equipment capacity of the solar collector and air source heat pump has decreased compared to scenario 1, but the capacity of photovoltaic modules and energy storage batteries is the highest among all scenarios. Compared with scenario 1, the annual carbon dioxide emissions have decreased by 120.35kg, and the annual system cost has decreased by 5.13%. Compared with Scenario 1 and Scenario 2, the capacity of equipment such as solar collectors, air source heat pumps, and photovoltaic modules in Scenario 3 has all decreased. Compared with Scenario 1, the annual carbon dioxide emissions have decreased by 28.11kg, and the annual system cost has decreased by 33.01%. In scenario 4, the area of the solar collector, the capacity of photovoltaic modules and energy storage batteries are the lowest among all scenarios. Compared with scenario 1, the annual carbon dioxide emissions are reduced by 120.35kg, and the annual system cost is reduced by 35.4%, resulting in the best regulation effect.

Upon integrating flexible electrical loads into regulation, the load schedule aligns with peak photovoltaic output periods, effectively utilizing solar power and thereby reducing the required capacities of energy storage batteries and diesel generators. When both flexible thermal and electrical loads are regulated concurrently, the capacities needed for photovoltaic modules, energy storage batteries, and diesel generators reach their minimum across all scenarios; After the participation of flexible heat load regulation, the indoor temperature decreases and the fluctuation amplitude is large. When flexible electricity and heat loads are simultaneously regulated, indoor temperature fluctuations tend to flatten.

5. Conclusion. The author proposes the use of digital twins and genetic algorithms for the study of building energy systems. The author analyzes the impact of flexible loads on the optimization design of multi energy complementary building energy systems and establishes a multi energy complementary building energy system optimization model and flexible load model in MATLAB/Simulk, mainly consisting of photovoltaic modules and solar collectors. The genetic algorithm is used to optimize the capacity of various equipment in the system, and the following conclusions are obtained:) Compared with not considering flexible loads, when flexible electricity loads, flexible heat loads, and flexible electricity/heat loads participate in regulation, the annual cost of the system is reduced by 5.24%, 33.11%, and 35.5%, respectively, and the annual carbon dioxide emissions are reduced by 120.46, 28.22, and 120.46kg, respectively. When flexible thermal and electrical loads are simultaneously regulated, the best reduction effect is achieved for the annual value of system costs and annual carbon dioxide emissions; After the participation of flexible electrical loads in regulation, the load shifts towards the photovoltaic output period, timely consuming photovoltaic power generation and reducing the capacity of energy storage batteries and diesel generators. When flexible thermal and electrical loads are simultaneously regulated, the capacity of photovoltaic modules, energy storage batteries, and diesel generators is the lowest value among all scenarios; After the participation of flexible heat load regulation, the indoor temperature decreases and the fluctuation amplitude is large. When flexible electricity and heat loads are simultaneously regulated, indoor temperature fluctuations tend to flatten.

REFERENCES

[1] Zhang, J. (2023). The effect of carbon tax incidence on household energy demand and welfare in the u.s. Environmental Science and Pollution Research, 30(5), 13210-13223.

[2] Li, M., Zhu, K., & Lu, Q. Y. K. (2023). Technical and economic analysis of multi-energy complementary systems for net-zero energy consumption combining wind, solar, hydrogen, geothermal, and storage energy. Energy conversion & management, 295(11), 1-17.

[3] Abdelmoumene, A., Bentarzi, H., Iqbal, A., & Krama, A. (2024). Developments and trends in emergency lighting systems: from energy-efficiency to zero electrical power consumption. Life Cycle Reliability and Safety Engineering, 13(2), 129-145.

[4] Tanriverdi, B., & Gedik, G. Z. (2023). Importance of hvac system selection in reducing the energy consumption of building retrofits–case study: office building in london. Civil Engineering and Architecture, 11(1), 217-227.

[5] Teng, J., Yin, H., & Wang, P. (2023). Study on the operation strategies and carbon emission of heating systems in the context of building energy conservation. Energy Science And Engineering, 11(7), 2421-2430.

[6] Roodkoly, S. H., Fard, Z. Q., Tahsildoost, M., Zomorodian, Z., & Karami, M. (2024). Development of a simulation-based ann framework for predicting energy consumption metrics: a case study of an office building. Energy efficiency, 17(1), 1-24.

[7] Hawks, M. A., & Cho, S. (2024). Review and analysis of current solutions and trends for zero energy building (zeb) thermal systems. Renewable & sustainable energy reviews, 189(Jan. Pt.B), 1-15.

[8] Altes-Buch, Q., Quoilin, S., & Lemort, V. (2022). A modeling framework for the integration of electrical and thermal energy systems in greenhouses. Building Simulation, 15(5), 779-797.

[9] Bazenkov, N. I., Dushin, S. V., & Mikhail V. GoubkoVsevolod O. KorepanovYuriy M. RassadinLeonid A. SeredaAlla G. Shinkaryuk. (2022). An office building power consumption dataset for energy grid analysis and control algorithms. ifac papersonline, 55(9), 111-116.

[10] Libralato, M., D'Agaro, P., & Cortella, G. (2023). Development of an energy digital twin from a hotel supervision system using building energy modelling. IOP Publishing Ltd, 53(Oct. Pt.C), 1-8.

[11] Ohmura, T., Shimomura, Y., Egawa, R., & Takizawa, H. (2023). Toward building adigital twin ofjob scheduling andpower management onanhpc system, 7(18), 4654-4667.

[12] Hou, Y., & Volk, R. (2022). Conceptual design of a digital twin-enabled building envelope energy audits and multi-fidelity simulation framework for a computationally explainable retrofit plan. Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 30(5), 1-22.

[13] Nobuyoshi, Y. J. R., Antonio, V. O. A., & Angelo Peixoto da Costa J.Suemy Arruda Michima P.de Novaes Pires Leite G.Claudius Nunes Silva H.Gabriel Carvalho de Oliveira E.Tiba C. (2023). Real-time energy and economic performance of the multi-zone photovoltaic-drive air conditioning system for an office building in a tropical climate. Energy conversion & management, 297(Dec.), 1-21.

[14] Cheraghi, R., & Hossein, J. M. (2023). Multi-objective optimization of a hybrid renewable energy system supplying a residential building using nsga-ii and mopso algorithms. Energy conversion & management, 294(Oct.), 1-20.

[15] Juan M. González-Caballín Sánchez, A. Meana-Fernández, J.C. Ríos-Fernández, & A.J. Gutiérrez Trashorras. (2023). Characterization of housing stock for energy retrofitting purposes in spain. Building Simulation, 16(6), 947-962.

[16] Saleh BabaaAbdul Aziz Al RawahiAngala SubramanianAbdullah Humaid AlshibliShahid KhanMartin KhzouzMuneer Ahmed-Ibrahim Ashrafi. (2022). Smart building design to improve the energy consumption at an office room, 13(9), 209-221.

[17] Wang, H., Ma, W., & Wang, Z. L. C. (2022). Multiscale convolutional recurrent neural network for residential building electricity consumption prediction. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 43(3), 3479-3491.

[18] Tahmasebi, M., & Nassif, N. (2022). An intelligent approach to develop, assess and optimize energy consumption models for air-cooled chillers using machine learning algorithms. American journal of engineering and applied sciences, 15(3), 220-229.

[19] Gholamian, E., Bagheri, B. R., & Zare, V. R. S. F. (2022). The effect of incorporating phase change materials in building envelope on reducing the cost and size of the integrated hybrid-solar energy system: an application of 3e dynamic simulation with reliability consideration. Sustainable Energy Technologies and Assessments, 52(Aug. Pt.B), 1-13.

[20] Aliabadi, A. A., Chen, X., & Yang, A. S. K. (2023). Retrofit optimization of building systems for future climates using an urban physics model. Building and environment, 243(Sep.), 1-20.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 24, 2024
Accepted: Aug 12, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1440–1447, DOI 10.12694/scpe.v26i3.4171

THE CONSTRUCTION OF MATHEMATICAL MODEL OF SWIMMERS' TECHNICAL MOVEMENTS USING MULTIMODAL DEEP LEARNING FRAMEWORK

MENGMENG WANG*AND YANGWEN HE†

Abstract. This paper proposes a mathematical model construction method based on a multi-modal deep learning framework aiming at the accuracy and real-time requirements of swimmers' technical movement analysis. The model can extract the image features and timing information of athletes' movements from video sequences by integrating spatiotemporal modules. This paper introduces the translation partial channel strategy to overcome the limitation of spatiotemporal information separation in traditional methods, which can seamlessly integrate spatiotemporal features and enhance the recognition ability of complex action patterns. In addition, NetVLAD is used as the feature aggregation layer. This layer can capture and encode the global and local features of the athlete's movements, thereby improving the classifier's performance. In the experimental part, the model is strictly verified, and the results show that compared with the prior art, the model in this paper shows higher accuracy and faster processing speed in the swimmer's action classification task. This provides the possibility of immediate feedback for coaches and athletes and lays a solid foundation for further research in the field of sports science.

Key words: Multimodal deep learning; Space-time module; Translation part of the channel; NetVLAD; Classification of swimmers.

1. Introduction. In pursuing excellence in sports competition, technical improvement and scientific analysis are the driving forces to promote athletes to reach their peak state. Swimming, an ancient and vibrant water sport, attracts millions of fans and professionals worldwide with its unique charm. With the progress of science and technology and the continuous innovation of data analysis methods, applying deep learning technology in swimming has gradually become essential to improve the training effect and competition performance. In particular, the emergence of multimodal deep learning frameworks has provided unprecedented opportunities to capture and analyze the complex technical movements of swimmers. The technical movement analysis of swimmers is a highly specialized work that requires accurate capture and in-depth understanding of multi-dimensional information such as the athlete's posture, power distribution and movement rhythm in the water. Traditional methods often rely on intuitive observation by experienced trainers or use expensive and cumbersome motion-capture systems. Although these methods can provide valuable information to some extent, their subjectivity, limitation and lack of real-time limit their wide application in practical training.

The rise of multimodal deep learning frameworks has brought new hope to solve this problem. This framework can comprehensively utilize various sensor data, such as video streams captured by high-speed cameras and time series data recorded by motion tracking devices, and automatically extract key features through advanced algorithms to build mathematical models that reflect the nature of athletes' technical movements. The spatiotemporal module plays a core role in this framework, which can simultaneously process the spatial structure of visual images and the temporal evolution of temporal information, providing a comprehensive and detailed data basis for motion analysis.

Behavior recognition is essential to behavior prediction, posture analysis, etc. Its core purpose is to realize the accurate cognition of pedestrian behavior characteristics in the video, and the most critical problem is to fully dig out the adequate information of various parts in the video. One of the main differences is how the timing information is used and modeled. The previous research mainly used spatial-time descriptors to extract and recognize image features. Literature [1] applies it to dense moving trajectory images. Reference [2] optimizes IDT and improves feature regularization and feature coding. Remarkable results have been achieved

*School of Physical Education, Jiangsu University of Technology, Changzhou 213001, Jiangsu, China (Corresponding author, 2022500025@jsut.edu.cn)

†Department of Orthopedics, Changzhou West the Taihu Lake Hospital, Changzhou 213149, Jiangsu, China

in motion recognition. In recent years, with the rise of deep learning technology, more and more research has been done on image features. The 2-D convolution model only extracts a specific image from a single frame in video understanding, and it cannot be effectively modeled as a time series. In reference [3], for the problem of human behavior feature extraction, it is proposed to use a two-layer parallel convolutional network for image deep learning and a two-channel convolutional neural network for human behavior extraction to identify human behavior efficiently. In literature [4], a time-sequence segmented network was established based on the two-stream network, and a complete time-sequence data set was designed through segmented training of videos. The 3D convolution algorithm can better capture the temporal and spatial information, but it requires a large amount of computation. In literature [5], a 3D convolutional neural network based on time-domain information is applied to behavior recognition for the first time, and a 3D kernel function is used to conduct feature extraction on both space-time and behavior dimensions. Literature [6] pooled a 3D convolutional neural network called C3D. With the increasing demand for real-time image processing technology, algorithms based on lightweight models have gradually become a research hotspot. Reference [7] establishes MFNet, a mobile feature network containing action modules—the effective fusion of spatial and temporal information between frames within the same frame. In literature [8], a simple and effective STM module is designed to encode space and motion information using a two-dimensional network as the framework. Literature [9] analyzes the dynamic content in videos by using the fusion of slow and high-resolution CNN and fast CNN, respectively. Two parallel convolutional neural networks are applied to the same video sequence to improve the image quality. Literature [10] introduces RGB and optical flow into the two-in-one first-line network. The motion information of the streaming image is obtained in the moving state layer to realize the precise control of the underlying RGB signal.

In this paper, an innovative translation partial channel method is proposed. Introducing translation operation in the feature extraction process makes the model more sensitive to capturing the subtle changes in the action process, thus enhancing the ability of spatiotemporal feature representation [11]. In addition, as an efficient feature aggregation technology, NetVLAD can reduce data redundancy while maintaining feature diversity, which provides strong support for action classification tasks. The research content of this paper focuses on the following key points. Firstly, this paper designs and implements a multi-modal deep learning framework with integrated spatiotemporal modules, which can automatically extract swimmers' technical motion features from continuous video streams. Secondly, utilizing the translation partial channel method, this paper optimizes the fusion process of spatiotemporal information and improves the feature representation ability of the model. The NetVLAD aggregation mechanism is used again to build a compact and efficient feature descriptor, which provides a solid foundation for the subsequent action classification [12]. Finally, through many experiments, this paper shows the superior performance of the proposed framework on the task of swimmer's movement classification.

2. Swimmers' technical movement recognition system.

2.1. System Architecture. This project begins with a systematic definition of swimming behavior, regarded as a set of actions on different time series. A complete human behavior database is formed through data collection, preprocessing and feature extraction. This project addresses global characteristics such as swimming speed, stroke frequency and lap time and specific characteristics such as intensity and duration of individual movements [13]. The overall design of the system is shown in Figure 2.1. A single pose sensor obtains the movement information of athletes. Using the swimming image collected by the high-speed camera, the corresponding actual behavior markers are extracted, and the swimming database is built by combining the feature quantity. The model is divided into two parts: the first is to classify and identify the motion behavior to maximize the utilization of the training sample; The second is to verify the model. A series of continuous behavior sequences are obtained by analyzing the motion information obtained by each sensor component.

2.2. Data acquisition device and experiment. In the test, a 36 mm× 51.3mm×21mm integrated pose sensor module was used to complete the measurement of swimming posture. The sensor has a three-way acceleration sensor, a three-dimensional gyroscope and a three-way geomagnetic field sensor. A high-performance microprocessor is used for acquisition. Kalman dynamic filter is used to obtain the real-time pose of the sensor component [14]. Through WIFI wireless transmission technology, realize the sensor components

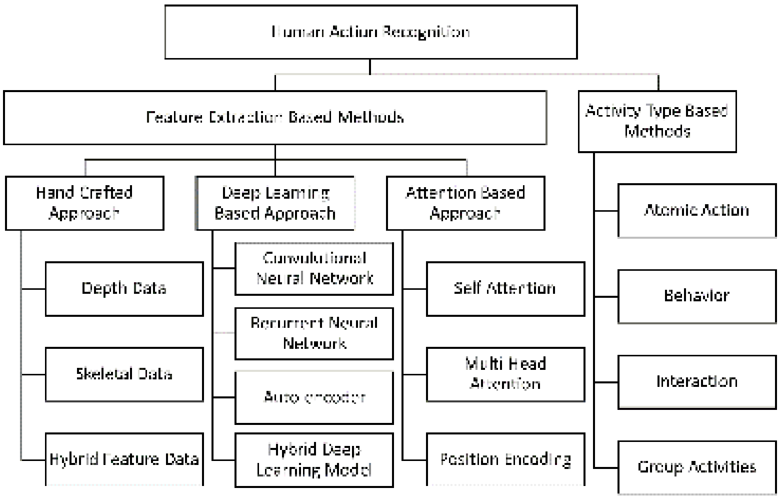


Fig. 2.1: Swim Action Recognizer system enclosure.

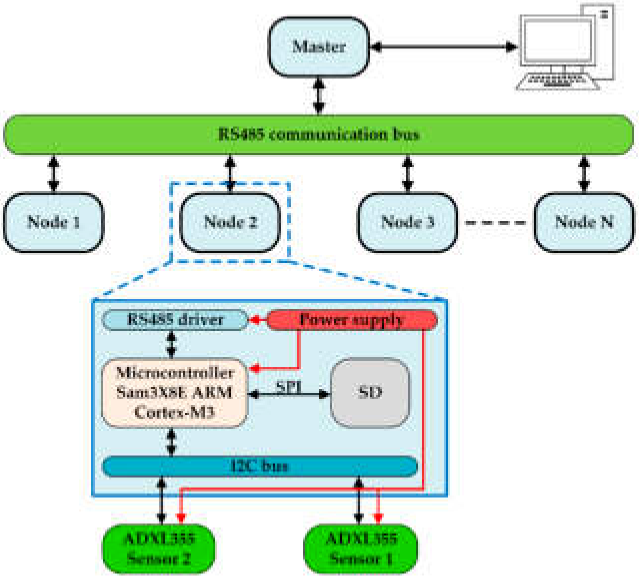


Fig. 2.2: Structure of swimmer condition monitoring sensor assembly.

of the pose and timely sequence information transmission. When the sensor component stops working for 5 minutes, it will automatically enter sleep mode, thus saving energy. Sleep automatically returns to its normal operating mode when the action is activated. The sensor is powered by 3.3V 5.0V. A USB charging port is installed externally. Figure 2.2 is the structure of swimmer condition monitoring sensor assembly. The sensor group's power supply is provided according to the swimming action characteristics. External USB charging port. Attach the sensor assembly to the waist's center using a strap (Figure 2.3). The position, acceleration, Angle and velocity sensor's arrow point to the square of each axis. Considering that different swimming conditions will lead to the change of geomagnetic field, this paper gives up the detection of geomagnetic field, and only uses two different motion states, such as acceleration and angular velocity [15]. The high-speed camera Q2m takes

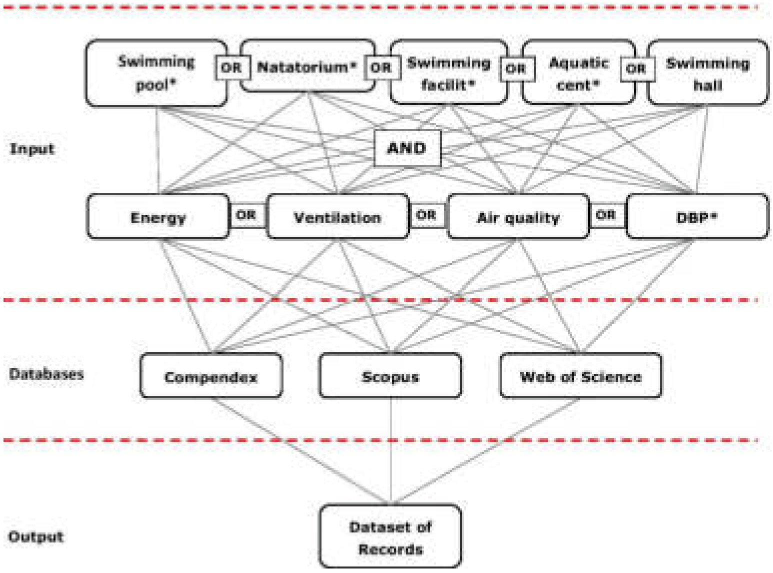


Fig. 2.3: Swimming data acquisition.

real-time photos of swimming motion marks at 5000 frames per second. The timing of an athlete's swimming movements can be watched frame by frame via video synced with sensor data.

2.3. Shift space-time module. An image sequence analysis method based on wavelet transform is proposed. In contrast to the convolution operation, the move operation does not require parameter values and floating-point operations but contains a set of operations with storage properties. Using 1x1 convolution for data fusion can reduce the computational cost. For example, in generic one-dimensional convolution, the prediction is expressed as a value derived from a weighted sum of the various inputs [16]. On the other hand, an input value is regarded as the input of the present moment and the adjacent moment. The input value is the input value of the three time points after displacement +1, 0, 1, multiplication, and addition. This kind of displacement convolution can be reduced to two processing methods: one is translation operation and the other is multiplication operation.

$$F_i = \lambda_1 T_{i-1} + \lambda_2 T_i + \lambda_3 T_{i+1} \tag{2.1}$$

$$T_i^{-1} = T_{i-1}, T_i^0 = T_i, T_i^{+1} = T_{i+1} \tag{2.2}$$

$$F = \lambda_1 T^{-1} + \lambda_2 T^0 + \lambda_3 T^{+1} \tag{2.3}$$

After sorting the Z-channel input in the K-frame picture, the tensor is shown in Figure 2.4.

Each image channel represents the amount of image frame features captured at each moment. The characteristic quantity is simultaneously transversally shifted for multiple channels along the time direction. Some channel values are down one space, and some are shifted one space. The blank part is filled with 0, and the excess channel value of the feature image is transferred out, thus completing the translation of the two directions [17]. After the movement, the characteristic information of the neighboring frame is merged with the current frame. But more movement doesn't mean more exchange of information. When the displacement is too small, the function of the timing model cannot satisfy the correlation of complex timing. When the displacement is large, the learning effect of spatial characteristics will decrease. By adjusting only the local channel, the efficient union of multiple channels is realized. The displacement model is to be added to the residual blocks

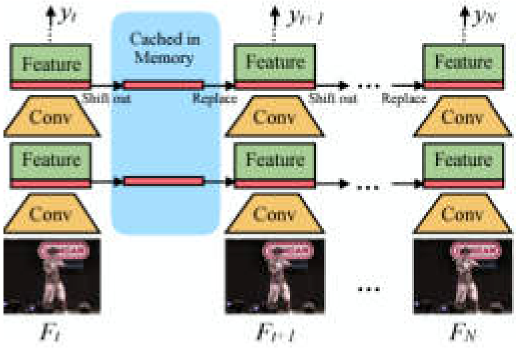


Fig. 2.4: Schematic diagram of the characteristics of the shift module.

of each branch of the residual network. Before the convolution operation, the displacement operation is carried out to complete the fusion of space-time spatial information without increasing the cost of three-dimensional computation. The time domain perception outfield value of each shift space unit is increased by 2 times to complete the construction of the time domain model.

2.3.1. Multi-mode. The multi-mode data processing of the image three-color difference is carried out based on fully mining the space and time information. In this way, the image is enhanced. Traditional feature-based image processing algorithms have a lot of operational overhead. This project intends to convert image color differences into RGB differences [18]. Then, the apparent changes and prominent moving areas are modeled to obtain the motion features of the image. Finally, the predicted value and the scores obtained from the spatial and temporal characteristics are added together to obtain the corresponding identification results.

2.3.2. NetVLAD Method. VLAD is based on locally aggregated information vectors, expressed through post-processing, and introduced into the terminal convolutional neural network to achieve image feature expression. In this paper, the NetVLAD algorithm is applied to the convolutional network and used as a pooling layer in the convolutional network to gather the characteristic information in the network. For A feature graph t, an S dimensional feature vector $t_i \in D^S$ must be obtained from the spatial position to represent the feature graph. First, J cluster centers z_j are given. The feature space D^s is divided into J units. Each feature vector t_i corresponds to a unit, and the residual vector $t_i - z_j$ is used to represent the difference between the feature vector and the cluster center. The resulting difference vector is expressed as:

The paper first needs to find its eigenvector A in the dimension S in space. First, J cluster centers B are given [19]. The model is divided into feature space C. Each feature vector D corresponds to the cell, and the residual vector E expresses the difference between the feature vector and the cluster center.

$$H(j,j) = \sum_{i=1}^{N} \frac{e^{-\beta\|t_i - z_j\|^2}}{\sum_{j'} e^{-\beta\|\|_i - z_j\|^2}} \left(t_i(j) - z_j(j)\right) \tag{2.4}$$

$t_i(j)$ and $z_j(j)$ represent the j component of the eigenvector t_i and cluster center z_j, respectively, where β is a trainable super parameter. The J column of the input matrix $h \in D^{JS}$ represents the eigenvectors gathered in the J cell, and then the matrix is normalized to a column and L_2 is normalized to a one-dimensional vector H to describe this property. Finally, the input data is fed into the fully connected layer for classification.

3. Simulation results and analysis.

3.1. Accurate rate and rate of image recognition. The efficiency and stability of the algorithm are verified by testing the accuracy and speed of swimming pose images of moving targets. In this experiment 1, the moving object is always within the shooting area of the camera. This paper counts the number of cooperative times and the time used [20]. The Kalman prediction and SIFT were combined to carry out the test, taking the swimmer's stroke as the research object (Figure 3.1). The calculation results of the optimal pairing rate

Fig. 3.1: Image recognition comparison diagram between traditional algorithm and the proposed algorithm.

Table 3.1: Comparison effect of moving objects.

Algorithm	Matches (N)	Successes (N)	Success rate (%)	Match time (S)	Average time per frame (S)
Textual algorithm	104	94	94	15	0.15
Traditional algorithm	104	68	68	197	1.97

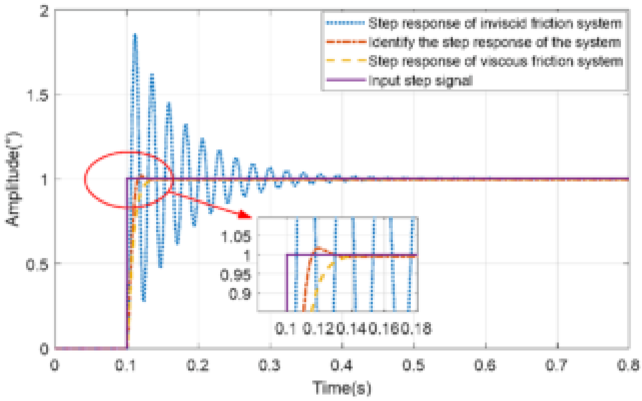


Fig. 3.2: Unit step response curve of the system.

and single frame time of the two algorithms are obtained (Table 3.1). This method can accurately classify the pose images of moving objects.

3.2. Image recognition effect. Experiments verify the effectiveness of this method, and it is tracked and identified. Experiments compare the performance of PID controller and fuzzy PID controller. The first is the Ziegler-Nichols algorithm. Then, the fuzzy PID controller is designed using the PID parameter set, and the step performance curve of the PID controller is recorded. The performance curve combining the above two controls uses the median filtering method (Figure 3.2).

The two systems' temperature rise and adjustment time are analyzed, and the results of related parameters

Table 3.2: Comparison table of step reaction diagram.

Controller	Overshoot	Rise time (t/s)	Adjustment time (t/s)
Conventional PID	0.25	2.08	4.17
Fuzzy PID	0.13	0.83	2.19

are obtained. The traditional PID method is used to adjust the attitude tracking of the moving object, and the lifting time is 2.08 seconds, the overshoot is 25%, and the adjustment time is 4.17 seconds. Using fuzzy PI D to adjust, the lifting time is reduced by 1.3 seconds. This reduces overshoot by 12%, reduces adjustment time by 1.9 seconds and improves the camera's performance in tracking moving objects. The algorithm in this paper completes the automatic recognition of moving objects and the tracking and retrieval of matching adjustment models (Table 3.2).

4. Conclusion. Using a multi-modal deep learning framework, this paper successfully constructs a mathematical model of swimmers' technical movements. By integrating a spatiotemporal module, this paper effectively extracts the image features and timing information of athletes' movements from video data, which provides a rich data basis for movement analysis. Applying the translation partial channel method further optimizes the fusion of spatio-temporal information, and enhances the recognition ability of the model for complex motion patterns. The introduction of the NetVLAD aggregation mechanism enables the model to process a large amount of feature information efficiently, significantly improving the accuracy of action classification. The experimental results show that this model performs excellently in classifying swimmers' movements, which provides a new technical analysis tool for coaches and athletes.

Acknowledgement. This work has been supported by a Social Science Project of Talent Introduction of Jiangsu University of Technology in 2022. The subject is Research on the Integration Development of National Traditional Sports and Early Childhood Physical Education Under the Background of Constructing a Sports Power, No.KYY22525.

REFERENCES

[1] Li, Z., Ye, X., & Liang, H. (2023). Sports video analysis system based on dynamic image analysis. Neural Computing and Applications, 35(6), 4409-4420.

[2] Strömbäck, D., Huang, S., & Radu, V. (2020). Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1-22.

[3] Chen, L., & Hu, D. (2023). An effective swimming stroke recognition system utilizing deep learning based on inertial measurement units. Advanced Robotics, 37(7), 467-479.

[4] Xia, H., Khan, M. A., Li, Z., & Zhou, M. (2022). Wearable robots for human underwater movement ability enhancement: A survey. IEEE/CAA Journal of Automatica Sinica, 9(6), 967-977.

[5] Zhang, X. (2021). Application of human motion recognition utilizing deep learning and smart wearable device in sports. International Journal of System Assurance Engineering and Management, 12(4), 835-843.

[6] Zhang, Y. (2023). Track and field training state analysis based on acceleration sensor and deep learning. Evolutionary Intelligence, 16(5), 1627-1636.

[7] Yang, M., & Zhang, S. (2023). Analysis of sports psychological obstacles based on mobile intelligent information system in the era of wireless communication. Wireless Networks, 29(8), 3599-3615.

[8] Amsaprabhaa, M. (2024). Hybrid optimized multimodal spatiotemporal feature fusion for vision-based sports activity recognition. Journal of Intelligent & Fuzzy Systems, 46(1), 1481-1501.

[9] Chinchilla Gutierrez, S., Salazar, J., & Hirata, Y. (2022). Mixed-reality human-machine-interface for motor learning of physical activities. Advanced Robotics, 36(12), 583-599.

[10] Kaseris, M., Kostavelis, I., & Malassiotis, S. (2024). A Comprehensive Survey on Deep Learning Methods in Human Activity Recognition. Machine Learning and Knowledge Extraction, 6(2), 842-876.

[11] Matsuyama, H., Aoki, S., Yonezawa, T., Hiroi, K., Kaji, K., & Kawaguchi, N. (2021). Deep learning for ballroom dance recognition: A temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing. IEEE sensors journal, 21(22), 25437-25448.

[12] McGrath, J., Neville, J., Stewart, T., & Cronin, J. (2021). Upper body activity classification using an inertial measurement unit in court and field-based sports: A systematic review. Proceedings of the institution of mechanical engineers, Part P: Journal of sports engineering and technology, 235(2), 83-95.

[13] Talha, M. (2022). Research on the use of 3D modeling and motion capture technologies for making sports training easier. Revista de Psicología del Deporte (Journal of Sport Psychology), 31(3), 1-10.

[14] Chen, G. (2024). An interpretable composite CNN and GRU for fine-grained martial arts motion modeling using big data analytics and machine learning. Soft Computing, 28(3), 2223-2243.

[15] Ramesh, M., & Mahesh, K. (2023). Efficient key frame extraction and hybrid wavelet convolutional manta ray foraging for sports video classification. The Imaging Science Journal, 71(8), 691-714.

[16] Van Leeuwen, T. (2021). The semiotics of movement and mobility. Multimodality & Society, 1(1), 97-118.

[17] Siddiqi, M. H., Alshammari, H., Ali, A., Alruwaili, M., Alhwaiti, Y., Alanazi, S., & Kamruzzaman, M. M. (2022). A template matching based feature extraction for activity recognition. CMC-COMPUTERS MATERIALS & CONTINUA, 72(1), 611-634.

[18] Kanwal, S., Khan, F., & Alamri, S. (2022). A multimodal deep learning infused with artificial algae algorithm–An architecture of advanced E-health system for cancer prognosis prediction. Journal of King Saud University-Computer and Information Sciences, 34(6), 2707-2719.

[19] Akila, K. (2022). Recognition of inter-class variation of human actions in sports video. Journal of Intelligent & Fuzzy Systems, 43(4), 5251-5262.

[20] Turmo Vidal, L., Márquez Segura, E., & Waern, A. (2023). Intercorporeal Biofeedback for Movement Learning. ACM Transactions on Computer-Human Interaction, 30(3), 1-40.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 25, 2024
Accepted: Jul 30, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1448–1456, DOI 10.12694/scpe.v26i3.4175

INTELLIGENT EVALUATION AND PREDICTION MODEL OF MENTAL HEALTH STATUS BASED ON DEEP LEARNING

BINGSAI CHEN*

Abstract. In order to solve the problems of low efficiency and accuracy in traditional social psychological measurement techniques, the author proposes a deep learning based intelligent evaluation and prediction model for mental health status. The author combines multi parameter acquisition technology with deep learning algorithms and designs a psychological crisis testing algorithm based on a bipartite graph convolutional network model, using graph convolutional networks as the foundation. The algorithm is then embedded with a psychological testing instrument. The experimental results show that the accuracy of the BGCN undirected model is 88.52%, the accuracy is 86.21%, the recall is 66.56%, and the F1 value is 61.20%, all of which have performance advantages in model comparison. At the same time, with the increase of iteration times, the Loss change curve shows a stable downward trajectory, while the accuracy and F1 value curves show a large fluctuation amplitude in the early stage, a small fluctuation amplitude in the later stage, a fast decline speed in the early stage, and a stable trend in the later stage. This indicates that the model has the ability to conduct stable and accurate testing in the later stage of iteration. From the experimental results, it can be seen that this model can perform accurate psychological testing, which is conducive to the active development of social psychological testing.

Key words: Multiple parameters, Deep learning, GCN,BGCN, Psychological testing, Graph convolution

1. Introduction.

Psychological health is an important component of health and also the foundation for the comprehensive development of individuals [1]. Scientific and efficient mental health assessment and intervention are prerequisites for effective psychological services [2]. However, traditional mental health assessments and interventions face significant challenges in terms of authenticity, effectiveness, and convenience when applied on a large scale [3].

With the rapid development of the current economy and society, people are constantly improving their quality of life while facing increasing pressure in their work, study, and life, which in turn has given rise to negative social phenomena and caused adverse social impacts [4]. At present, the application of mental health mainly revolves around physiological transaction detection and processing, including sign data detection and analysis, remote medical diagnosis services, real-time mobile ward monitoring, etc. However, there are relatively few applications in personal mental health testing, psychological counseling, and mental health message push [5]. The traditional methods of mental health assessment are mostly conducted through questionnaire surveys or face-to-face conversations, which have many shortcomings. At the same time, the accuracy of identifying individuals with psychological crises is seriously insufficient, and manpower is needed to supplement and distinguish the test results. Traditional methods of mental health assessment rely on the consultation and questionnaire surveys of doctors or psychological counselors. The diagnostic results generally depend on the experience of psychological researchers and the honesty of testers, and are easily affected by subjective differences, which may lead to misdiagnosis, missed diagnosis, inconsistent diagnosis before and after. In recent years, with the rapid development of artificial intelligence and big data technology, researchers have been able to more easily obtain richer multimodal data (such as speech data, text data, physiological data, etc.), and have also begun to try and use machine learning, deep learning and other methods in the field of artificial intelligence to characterize and model the relationship between these high-dimensional, unstructured, naturally generated data and their psychological state, achieve intelligent evaluation of psychological health status, and upgrade and replace psychological health intervention methods [6].

*School of Automotive Engineering, Henan Industry and Trade Vocational College, Zhengzhou, Henan, 450012, China. (Corresponding author, 13607693901@163.com)

2. Literature Review. Psychological state is not only a subjective feeling of an individual, but also an objective state that exists. Individuals with normal mental and psychological states and individuals with abnormal mental and psychological states have differences in their corresponding physiological parameters. By collecting physiological signals, these physiological differences can be detected and analyzed to determine whether an individual is in an abnormal state. Changes in mental and psychological states can cause changes in physiological signals, such as an accelerated heartbeat when a person is fearful. Physiological signals, due to their properties that are not easily concealed, can more objectively reflect the true mental state and psychological feelings. Monitoring changes in physiological signals has great practical significance in analyzing mental health problems. Therefore, identifying mental health problems based on physiological signals has gradually become a hot topic in the current field of mental health assessment research. Intelligent assisted diagnosis has been studied in various fields of mental health, classifying and diagnosing various psychological disorders, including anxiety, schizophrenia, depression, autism spectrum disorder, or attention deficit hyperactivity disorder. Lee, M. et al. analyzed the House Tree People Test (HTP), a widely used psychological test for drawing in clinical practice, which utilizes object detection techniques to extract more diverse information from images [7]. Zou, C. and others believe that the convenience of big data processing technology has played a huge advantage in many scenarios, and its deep learning can effectively mine different types of data in the dataset. Applying this method to the mining of psychological prediction datasets for legal misconduct can effectively prevent the occurrence of illegal behavior. Effective analysis of their psychological characteristics and emotional changes may pose hidden dangers, therefore it is necessary to extract such data in such situations [8]. Meng, Q. and others applied neural network algorithms of Bi LSTM and CNN models to study text data, and ultimately achieved high accuracy in psychological analysis experiments, providing a feasible solution for batch rapid analysis of psychological changes reflected in daily texts of basketball players [9].

Therefore, computer technology can be used to design a psychological testing instrument for individuals with sub healthy status in the social center. The graph convolutional model itself has high generalization ability and is more accurate in identifying and classifying information feature patterns. It can accurately classify structural information in model nodes and is very suitable for application in the field of psychological testing. Starting from this perspective, the research aims to design a psychological testing instrument based on multi parameter acquisition and bipartite graph convolutional neural network (BGCN), providing a practical approach for intelligent psychological testing.

3. Method.

3.1. Psychological testing method based on GCN algorithm.

The author first uses Graph Convolutional Network (GCN) to represent the vectors of subjects by transmitting information similar to those between subjects, and based on this, identifies the psychological status of the subjects [10]. Unlike traditional machine learning models, the construction of GCN is based on graph structured data, so it requires an appropriate graph structure; At the same time, GCN also requires data features from traditional machine learning, therefore, it should consider the generation of node features in the graph structure. The psychological state testing framework based on GCN includes four parts: data layer, preprocessing layer, model construction layer, and prediction layer, as shown in Figure 3.1.

In Figure 3.1, the data layer collects data and parameters for psychological testing from different data sources, including evaluations of the psychological status of the subjects and records of psychological disorders; The preprocessing layer preprocesses the collected data, extracts the characteristics of the subjects from the processed data, and constructs a psychological similarity map of the subjects based on their psychological states. The characteristics of the subjects and the psychological similarity map of the subjects serve as the basis for constructing the GCN model; The model construction layer constructs a GCN model using training samples based on preprocessed subject characteristics and subject psychological similarity maps; The prediction layer uses the trained GCN model to identify and predict participants in psychological tests [11].

Regarding the relationship between subjects, the author constructs a psychological similarity map based on their psychological state. Set the psychological similarity between subject a and subject b to $S_{a,b}$, and set the threshold to ω. If the value of $S_{a,b}$ exceeds the threshold ω, it is considered that the psychological states of the two subjects are similar. Therefore, a relationship line is added between the two subjects. The author uses cosine similarity as the criterion for judging the psychological similarity and interval of the subjects. If the

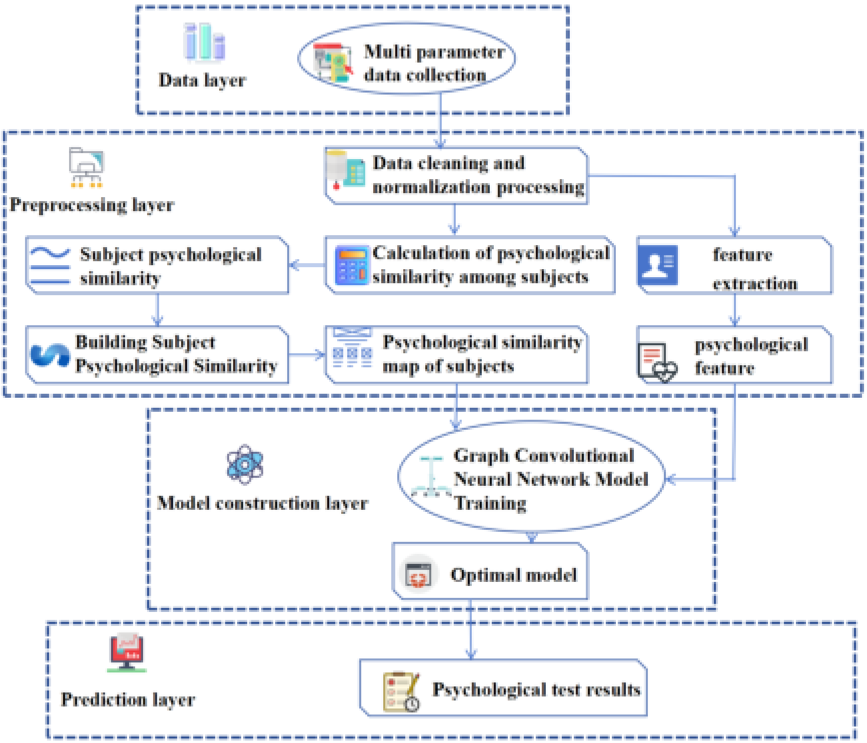

Data layer
Multi parameter data collection
Preprocessing layer
Data cleaning and normalization processing
Subject psychological similarity
Calculation of psychological similarity among subjects
feature extraction
Building Subject Psychological Similarity
Psychological similarity map of subjects
psychological feature
Model construction layer
Graph Convolutional Neural Network Model Training
Optimal model
Prediction layer
Psychological test results

Fig. 3.1: GCN based psychological state testing framework

cosine value of two vectors is 0, it is considered that the two vectors intersect; When the cosine value of two vectors is greater than 0, it is considered that they are somewhat similar. The psychological similarity graph of the subjects is a symmetric undirected graph, which is an image where the edges between two points do not have their own directional definitions and need to be defined through cyclic variables [12]. The construction process is as follows: first, clean the collected psychological state data of the subjects to ensure the credibility and effectiveness of the data, and then use the Z-score normalization method to normalize the cleaned data, as shown in equation 3.1.

$$x' = \frac{x - \overline{x}}{\sigma} \tag{3.1}$$

In equation 3.1, x is the original data; x' is the result of normalization and follows a standard normal distribution; $\overline{x}$ is the average value of the original data; σ represents the standard deviation of the original data. After normalization, the psychological feature matrix of the subjects is obtained, and the cosine similarity between the sample data is calculated as shown in equation 3.2.

$$S_{a,b} \frac{X_a \cdot X_b}{||X_a|| \cdot ||X_b||} \tag{3.2}$$

In equation 3.2, $S_{a,b}$ represents the psychological similarity between subject a and subject b; X is the psychological characteristic matrix; X_a and X_b are the representation vectors of subject nodes S_a and S_b, respectively. When the psychological similarity between two nodes exceeds the threshold ω, it is considered that there is a certain degree of similarity between the two, and a psychological similarity graph T (S, U) is obtained. The elements in the i-th row and j-th column of its adjacency matrix M are calculated as shown in

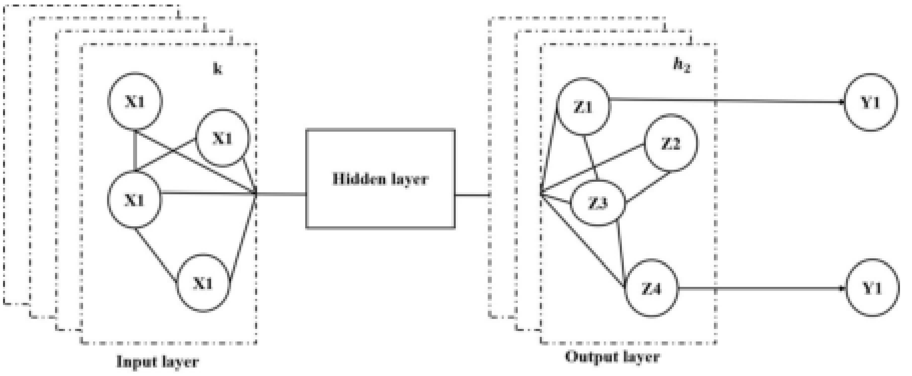


Fig. 3.2: Schematic diagram of model structure

equation 3.3.

$$M_{i,j} = \begin{cases} S_{a,b}, S_{a,b} > \omega \\ 0, S_{a,b} \leqslant \omega \end{cases} \tag{3.3}$$

The GCN model is abstractly represented as f (X, M), and its antecedent propagation is shown in equation 3.4.

$$G^{l+1} = \delta(\tilde{C}\tilde{C}^{-\frac{1}{2}}\tilde{M}\tilde{C}^{-\frac{1}{2}}G^l W^l) \tag{3.4}$$

In equation 3.4, H^l is the hidden feature matrix of the th layer; W^l is the weight matrix of the th layer neural network; δ is the activation function; $\tilde{M} = M + I_N$ is the new adjacency matrix, I_N is the N-dimensional identity matrix, and N is the number of subjects. The schematic diagram of the model structure is shown in Figure 3.2.

In Figure 3.2, k represents the dimension of the input feature vector X_i; h_2 represents the hidden feature vector Z_i dimension of the final output, and the result Y_i is obtained through Softmax classification. By calculating matrix $M = \tilde{C}^{-\frac{1}{2}}M\hat{C}^{-\frac{1}{2}}$ in advance, the forward propagation formula of the two-layer graph convolution model can be obtained as shown in equation 3.5.

$$f(X,M) = softmax(MReLU(MXW^{(0)})W^{(1)}) \tag{3.5}$$

The author chooses the ReLU function as the activation function, and its formula is shown in equation 3.6.

$$ReLU(x) = \begin{cases} 0, x \leqslant 0 \\ x, x > 0 \end{cases} \tag{3.6}$$

3.2. Design of a psychological testing instrument based on multi parameter acquisition and BGCN . In order to improve the testing accuracy and generalization of the model, the author set the tested individuals and testing metrics as two types of nodes in the convolutional network, with pairwise connections between the nodes forming a bipartite graph. A psychological testing method based on BGCN was designed. The two types of nodes have different properties, but they can be connected to each other, thereby reducing the coupling between nodes and ultimately improving the generalization and accuracy of the model. The architecture of BGCN based psychological testing method is similar to GCN, with the main difference being the construction of graphs and the training and testing of models. Assuming the set of indicators for psychological testing is P, the set of nodes for the subjects is R, and the set of edges in the bipartite graph formed by connecting the two is O, with the corresponding set of edge weights being W [13,14]. From this, the bipartite graph $T'(P,R,O,W)$ can be obtained as shown in Figure 3.3.

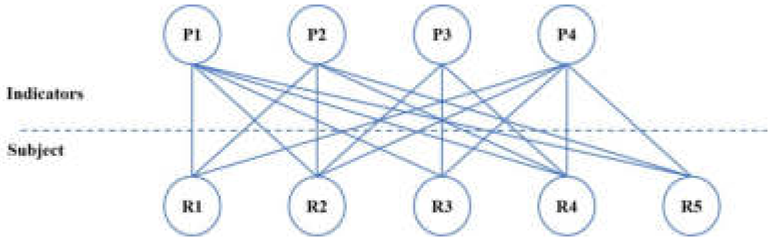


Fig. 3.3: Bipartite diagram

In Figure 3.3, subject node R_i and indicator node P_j are connected to each other, and the value of edge weight W is set based on the test indicator. Convert the judgment results obtained based on psychological testing indicators into numerical values. If there are n indicators that meet the degree of conformity, use the integer of the interval $[0, n-1]$ to represent the psychological status of the subjects based on the degree of conformity of the indicators. In practical situations, the subject bipartite chart cannot directly incorporate the transformed weight matrix, so it is necessary to standardize the weights of each indicator uniformly. There are two ways for the author to construct a bipartite graph. For an undirected graph $T_u(P, R, O, W)$, the edge weight set W directly corresponds to the values in the standardized weight matrix M; For directed graph $T_d(P, R, O, W)$, the edge weights of node R pointing to node P are set to equal values, and the edge weights of node P pointing to node R are set to the processed weight matrix M'. Assuming that the hidden features of the subject nodes are influenced by the indicator nodes, the degree of influence depends on the indicator matrix M of the subject nodes. When the influence of two types of nodes on each other is the same, an undirected graph is used. When the influence of each subject node on the indicator node is the same, a directed graph is used [15].

Assuming the indicator feature matrix is Q, $Q = I_D, I_D$ represents the D-dimensio- nal identity matrix; The subject feature matrix is E, which is a zero matrix of D and N, and the model feature matrix are horizontally connected; The bipartite graph convolutional neural network is represented as . The forward propagation formula for defining convolutional layers in bipartite graphs is shown in equation 3.7

$$\begin{cases} Q_k^{(l+1)} = \delta(\sum_{n \notin N_k} \alpha_{nk} E_n^{(l)} W^{(l)}) \\ E_n^{(l+1)} = \delta(\sum_{k \notin N_k} \alpha_{Kn} Q_k^{(l)} W^{(l)}) \\ X^{(l+1)}{=}[Q^{(l+1)} || E^{(l+1)}] \end{cases} \tag{3.7}$$

In equation 3.7, $Q^{(l)}$ represents the hidden feature of the l-layer indicator node; $E^{(l)}$ is the hidden feature of the -layer subject node; $a_{i,j}$ is the degree to which node j is affected by node i; $X^{(1)}$ is the hidden feature of layer l; $W^{(1)}$ is the neural network weight of the B-layer graph convolutional layer; N_i is the set of adjacent nodes i of a node. Unlike traditional convolutional neural networks, bipartite graph convolution requires two convolutions to transmit the information of the subject node back to that node. Therefore, the number of convolutions required to achieve node representation in this graph convolution method must be even. The two-layer graph convolution process is shown in Figure 3.4.

In Figure 3.4, on the l-th layer of graph convolution, the subject node first transmits feature information to the indicator node; On the l+1 layer graph convolution, the indicator node transmits feature information to the subject node. When training a model, if there are more parameters and fewer training samples, the model is prone to overfitting. In response to this phenomenon, the author added a Dropout layer after convolution of each layer of the graph. The Dropout layer improves model performance by preventing information transmission between hidden layer neuron nodes, and its information cannot be transmitted to the next layer. Then use the Softmax function to convert it into the classification probability of the subject, and use the negative logarithmic likelihood function as the loss function of the model. A psychological testing system based on multi parameter collection and deep learning should pay attention to the diverse needs of different subjects, as well as the common experience of the majority of subjects in their usage habits. Based on this principle, relevant techniques should

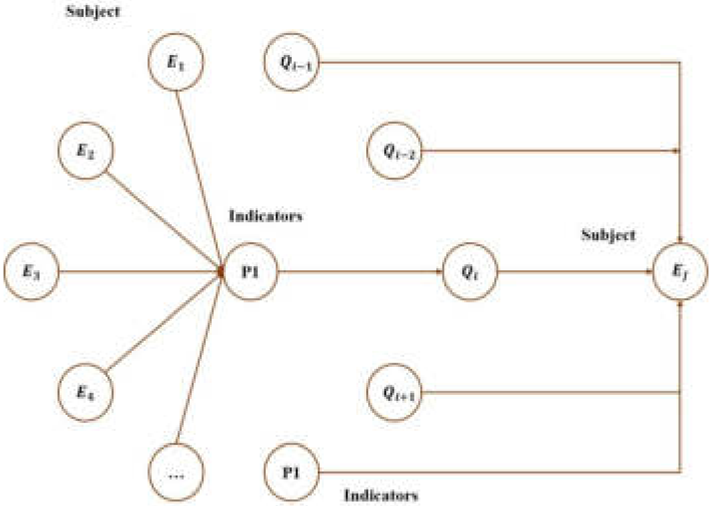


Fig. 3.4: Two layer graph convolution process

be used to set up the program layout. The author chose Myeclipse6.0 as the integrated development environment, MYSQL database, and J2EE as the network service environment. The actual operation of a psychological tester includes three parts: multi parameter collection, system analysis, and providing test results [16,18].

4. Results and Discussion. The author proposes a psychological crisis testing algorithm based on a bipartite graph convolutional network model, and designs a chimeric application combined with a psychological tester. In the process of performance analysis, the author selected an experimental sample set containing 260 positive samples and 1000 negative samples. 20% of the selected samples were used as the test set, and the remaining 80% were used as the training set. At the same time, the model sets the dropout probability to 0.5, the number of hidden features to 80 and 30, and the learning rate to 1c-2. The training process of the model first requires parameter initialization, setting initial parameters for each layer of BGCN. Existing initialization methods, such as Xavier initialization, can be used. Then conduct model training, calculate the predicted values through forward propagation, and use appropriate loss functions (such as cross entropy loss) to measure the difference between the predicted values and the true values. On this basis, a gradient descent optimizer is used to update the model parameters to minimize the loss function [19]. The author mainly uses the method of effect comparison for analysis. The author selects four main indicators: Accuracy, precision, recall, and F1 value, and analyzes them from three dimensions: composition threshold, number of training rounds, and horizontal comparison of model performance. The parameter dimension analysis is shown in Figure 4.1.

From Figure 4.1, it can be seen that in terms of accuracy, when the composition thresholds are 0, 0.1, and 0.2, the accuracy of the model is 88.52%, 84.27%, and 82.14%, respectively. It can be seen that as the composition threshold increases, the accuracy of the model gradually decreases; In terms of accuracy, when the composition thresholds are 0, 0.1, and 0.2, the accuracy of the model is 86.21%, 82.34%, and 80.96%, respectively. It can be seen that as the composition threshold increases, the accuracy of the model also decreases continuously; In terms of recall, when the composition thresholds are 0, 0.1, and 0.2, the accuracy of the model is 55.25%, 62.48%, and 58.52%, respectively. It can be seen that there is a peak change in recall, with an increase in recall in the early stage and a decrease in recall in the later stage. The optimal value point is located at the composition threshold of 0.1; On the F1 value, when the composition thresholds are 0, 0.1, and 0.2, the F1 values of the model are 0.6679, 0.6335, and 0.6123, respectively. It can be seen that as the composition threshold increases, the F1 value of the model continuously decreases. Overall, an increase in composition threshold will lead to a decrease in model performance [20].

In practical psychological testing applications, it is normal to form a testing period, and the stability of the model after the testing period is very important. Therefore, the model designed by the author is very suitable

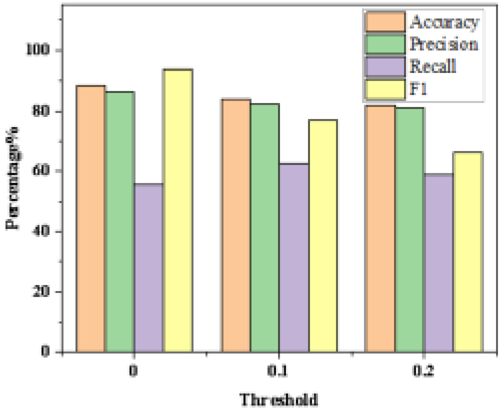


Fig. 4.1: Parameter dimension analysis

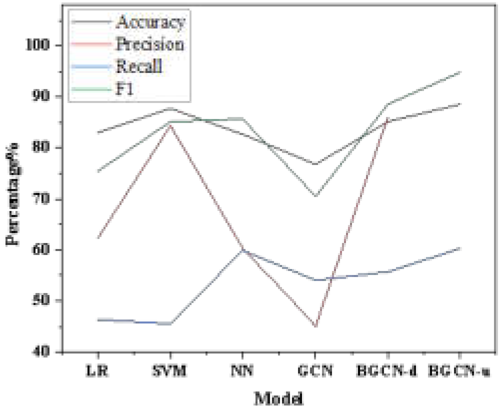


Fig. 4.2: Model Comparison Analysis

for application in psychological testing. The comparative analysis of the models is shown in Figure 4.2.

From Figure 4.2, it can be seen that in terms of accuracy, the BGCN undirected model designed by the author has an accuracy of 88.52%, the BGCN directed model has an accuracy of 85.18%, and the GCN model has an accuracy of 76.21%. The model designed by the author has the highest accuracy, meanwhile, compared with traditional machine learning algorithms of the same type, the model designed by the author also has a significant accuracy advantage; In terms of accuracy, the BGCN undirected model designed by the author has an accuracy of 86.21%, the BGCN directed model has an accuracy of 65.10%, and the GCN model has an accuracy of 44.46%. Compared with traditional machine learning algorithms of the same type, the model designed by the author also has the highest accuracy; In terms of recall, the BGCN undirected model designed by the author has a recall rate of 66.56%, the BGCN directed model has a recall rate of 54.72%, and the GCN model has a recall rate of 54.61%. At the same time, compared with traditional machine learning algorithms of the same type, the designed model has the highest recall rate; In terms of F1 value, the BGCN undirected model designed by the author has an F1 value of 61.20%, the BGCN directed model has an F1 value of 55.14%,

Table 4.1: Comparative Analysis Results

Data samples	The output value of the author's method	Expert evaluation value	Operation time/s	Assessment Level
1	9.57	9.7	1.82	good
2	6.46	6.4	1.64	preferably
3	4.56	4.6	1.67	commonly
4	7.45	7.4	1.51	good
5	5.82	5.8	1.58	preferably
6	2.36	2.3	1.44	difference
7	9.16	9.2	1.70	good
8	8.05	8.0	1.88	good
9	4.36	4.3	1.73	commonly
10	9.62	9.5	2.07	good
11	5.68	5.6	1.81	preferably
12	1.03	1.0	1.87	difference
13	7.55	7.6	1.64	good
14	8.56	8.6	1.77	good
15	8.24	8.2	1.40	good

and the GCN model has an F1 value of 53.71 %, the F1 value of the model designed for comparison with traditional machine learning algorithms of the same type is also the highest. From this, it can be seen that the performance of the model designed by the author is superior in terms of accuracy, precision, recall, and F1 value. When applied in psychological testers, more accurate psychological test results can be obtained, which helps to achieve intelligent psychological state detection.

The comparative analysis of the evaluation results obtained by this method and the expert evaluation results is shown in Table 4.1. According to Table 4.1, it can be seen that the difference between the psychological health evaluation values of college students output by this method and the expert evaluation values is very small, which can effectively obtain the psychological health evaluation results of college students and has a fast calculation speed.

5. Conclusion. The author proposes a research on an intelligent evaluation and prediction model for mental health status based on deep learning. Nowadays, the number of mental illness patients is increasing day by day, and mental health has become an important issue in current social research. The author designed a psychological crisis testing algorithm for a bipartite graph convolutional network model, which is attributed to graph convolutional technology and multi parameter acquisition technology, to address the relatively low accuracy and efficiency of traditional psychological measurement techniques at the social level. The algorithm optimized the model from two aspects: accuracy and generalization. The research results show that when the composition thresholds are 0, 0.1, and 0.2, the model accuracy, precision, recall, and F1 value all show a gradually decreasing trend with the increase of the composition threshold. As the number of iterations increases, the Loss value variation curve of the model designed by the author shows a relatively stable downward trend, while the accuracy and F1 value variation curve shows a relatively fast upward trend in the early stage. At the same time, the fluctuation range is large in the upward range, indicating that most of the erroneous cases occur in this range. In the later stage, the overall change trajectory tends to be stable, and the fluctuation is significantly reduced, indicating that the model can conduct high accuracy testing with stability at this time. In addition, in algorithm comparison, the BGCN undirected model designed by the author has an accuracy rate of 88.63%, an accuracy rate of 86.31%, a recall rate of 66.67%, and an F1 value of 61.30%, which is the highest performance in both improved models of the same type and traditional models. It can be seen from this that the model designed by the author has superior performance and can be used for more reliable and accurate psychological testing.

REFERENCES

[1] Ehsani, G. M., Namanya, W., Hughes, P., Frogh, A. K., & Safari, M. H. (2024). Six month evaluation of mental health, psychosocial support (mhpss) hotline of action contre la faim (acf), afghanistan. Intervention, 22(1), 11-16.

[2] Usui, K., Hasegawa, C., Ichihashi, K., Morita, K., Kano, Y., & Kanehara, A., et al. (2023). Development of a recovery-oriented early support program for adolescents and young adults with mental health problems. Japanese Journal of Brief Psychotherapy, 31(2), 37-48.

[3] Zhu, J. (2023). Mindfulness,smile,choice,practice:thoughts based on improving the mental health of adolescent secondary vocational students. Journal of Contemporary Educational Research, 7(11), 24-32.

[4] Kumar, A., Khanuja, K., Greene, N., Goudy, F., Green, A., & Gerolamo, A. (2024). Mental health diagnoses on the mini international psychiatric interview are associated with higher scores on the edinburgh postnatal depression scale. Nursing for Women's Health, 28(3), 177-186.

[5] Carlsson, L., Thain, E., Gillies, B., & Metcalfe, K. (2022). Psychological and health behaviour outcomes following multi-gene panel testing for hereditary breast and ovarian cancer risk: a mini-review of the literature. Hereditary Cancer in Clinical Practice, 20(1), 1-13.

[6] Sano, T., & Hamano, Y. (2023). An attempt to grasp the psychological state of athletes using a psychological testing system in college sports club activities. Japanese Journal of Sport Management, 15(1), 23-36.

[7] Lee, M., Kim, Y., & Kim, Y. K. (2024). Generating psychological analysis tables for children's drawings using deep learning. Data & knowledge engineering(Jan.), 149.

[8] Zou, C. (2022). The construction of psychological intervention mechanism of deep learning in the prevention of legal anomie. Frontiers in psychology, 13, 937268.

[9] Meng, Q. (2022). Psychological analysis of athletes during basketball games from the perspective of deep learning. Mobile Information Systems, 15(1), 23-36.

[10] Siritzky, M., Condon, D., & Weston, S. (2022). The role of personality in shaping pandemic response: systemic sociopolitical factors drive country differences:. Social Psychological and Personality Science, 13(1), 246-263.

[11] Proeschold-Bell, R. J., Stringfield, B., Yao, J., Choi, J., Eagle, D., & Hybels, C. F., et al. (2022). Changes in sabbath-keeping and mental health over time: evaluation findings from the sabbath living study:. Journal of Psychology and Theology, 50(2), 123-138.

[12] Chen, H., & Wu, H. (2023). Letter to the editor on "pain catastrophizing and pre-operative psychological state are predictive of chronic pain after joint arthroplasty of the hip, knee or shoulder: results of a prospective, comparative study at one year follow-up". International orthopaedics, 47(1), 281-282.

[13] Ueda, K., & Takada, S. (2022). Return and settlement processes and psychological changes in young people due to negative factors: a case study in unnan city, shimane prefecture. Journal of Rural Problems, 58(2), 59-66.

[14] Yang, M., Sheng, X., Ge, M., Zhang, L., Huang, C., & Cui, S., et al. (2022). Childhood trauma and psychological sub-health among chinese adolescents: the mediating effect of internet addiction. BMC psychiatry, 22(1), 762.

[15] Liu, Y., Liu, Y., Cheng, J., Pang, L. J., & Zhang, X. L. (2023). Correlation analysis of mental health conditions and personality of patients with alcohol addiction. World Journal of Psychiatry, 13(11), 893-902.

[16] Bah, F., & Kagotho, N. (2023). "if i don't do it, no one else will" narratives on the well-being of sub-saharan african immigrant daughters. Affilia, 39(2), 229-244.

[17] Ling, J., Lan, Y., Huang, X., & Yang, X. (2024). A multi-scale residual graph convolution network with hierarchical attention for predicting traffic flow in urban mobility. Complex & Intelligent Systems, 10(3), 3305-3317.

[18] Chen, L., Liu, R., Yang, X., Zhou, D., Zhang, Q., & Wei, X. (2022). Sttg-net:a spatio-temporal network for human motion prediction based on transformer and graph convolution network, 5(1), 224-238.

[19] Uchida, S., & Ikeda, H. (2022). The effect of working support using self-monitoring on the psychological state of patients with schizophrenia: through the daily report of the web system. Japanese Journal of Behavioral and Cognitive Therapies, 48(3), 261-271.

[20] Woodley, o. M. M. A., & Peaherrera-Aguirre, M. (2022). General intelligence as a major source of cognitive variation among individuals of three species of lemur, uniting g with g. Evolutionary Psychological Science, 8(3), 241-253.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jun 25, 2024
Accepted: Jul 31, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1457–1465, DOI 10.12694/scpe.v26i3.4235

THE APPLICATION OF IMAGE RECOGNITION TECHNOLOGY BASED ON DEEP LEARNING IN DATA ANALYSIS

WEI SHI[*] KAI GUO[†] WEILAN LIU [‡] AND JINGWEI GUO[§]

Abstract. To address the challenge of achieving high recognition accuracy across various image types, the author advocates for applying deep learning-based image recognition technology in data analysis research. The author first uses convolutional neural networks to train and process the raw laser image big data, extract image features, and set a threshold for pre segmentation to complete image preprocessing; then use the regularized least squares method to complete the laser image pattern recognition process and achieve image pattern differentiation; finally, construct an experimental section and analyze the application effect of this method. The experimental results show that the recognition rate of the preset target images for different types of images is stable at over 96%, the recognition error rate is stable at less than 2%, and the image recognition time is within 15 seconds, indicating that the method has good application effects. This method has a shorter recognition time and higher efficiency, providing impetus for the improvement of laser image analysis and processing technology.

Key words: Pattern recognition, Image preprocessing, Image analysis technology, Laser imaging, Spectral imaging technology, principal component analysis

1. Introduction. In the field of artificial intelligence research, intelligent image recognition technology is an emerging direction. This technology mainly focuses on analyzing various types of images as research objects. Due to the unique characteristics of different images, it is not possible to simply convert them into standardized image data [1]. Workers must process these images and then convert them into complex image data. On this basis, artificial intelligence technology is used to preprocess the data, select features based on the characteristics of the data, and select corresponding template matching models. Afterwards, massive data is classified using artificial intelligence and big data technology, and appropriate models are selected based on the analysis results. However, due to the many challenges involved in image generation, workers may encounter various problems during the image recognition process. The design work of an intelligent image recognition system is very complex, involving multiple different fields. Therefore, how to conduct information exchange is particularly important [2-3].

The quality and completeness of images play a crucial role in the accuracy of image recognition. Images with unclear or missing information pose challenges for accurate recognition. The integration of big data analysis and intelligent image recognition technologies offers a viable solution to enhance recognition accuracy. This fusion method involves optimizing database architecture and security design alongside advanced image recognition techniques. By adjusting pixel density based on image size and clarity, big data analysis ensures that data information integrates seamlessly into images. This approach alleviates the burden on data managers and simplifies the complexity of data management tasks. To address challenges in image recognition technology, it's crucial for stakeholders to prioritize technology integration and development [4-5]. By aligning with contemporary trends and leveraging advanced scientific concepts, innovations in image recognition can be guided and nurtured. The concept of big data involves processing vast amounts of data, and integrating this information into images facilitates user access. However, variations in image processing technology can lead to issues such as blurred images or poor pixel quality, limiting effective scanning or recognition capabilities [6]. To tackle challenges in scanning data images, personnel can harness the capabilities of domain adaptive

[*] Henan University of Science and Technology, Henan, Luoyang, 471000, China; LuoYang Polytechnic, Henan, Luoyang, 471000, China.

[†] Henan University of Science and Technology, Henan, Luoyang, 471000, China. (Corresponding author, guokai@haust.edu.cn)

[‡] LuoYang Polytechnic, Henan, Luoyang, 471000, China.

[§] LuoYang Polytechnic, Henan, Luoyang, 471000, China.

technology within big data sets. This technology enables proactive adjustments to image properties, including pixel density and clarity. It intelligently selects optimal pixel points within specified ranges based on image size. Big data analysis technology is pivotal in enhancing intelligent image recognition by accelerating recognition times, improving recognition quality and efficiency, and easing the workload of personnel [7].

2. Literature Review. Human perception and information acquisition of external things mainly rely on vision, and images are the description and recording of objective things by humans, collecting a large amount of information that cannot be described in words through images [8]. With the continuous updating and optimization of image acquisition technology, laser images have become the main type of image acquisition currently. Currently, multiple scholars have conducted targeted research in this field. Advanced deep learning algorithms offer promising solutions by converting non-image machine learning (ML) challenges into problems that can be tackled through image recognition techniques. Kovalerchuk, B. et al. introduced the CPC-R algorithm, which transforms non-image data into visual representations akin to images. Subsequently, deep learning CNN algorithms are employed to address learning tasks based on these visualized representations. This method has targeted effects on facial and handwritten digit data, but it does not have corresponding advantages for other types of image recognition and does not have wide applicability [9]. Rani, P. et al. undertook a study focused on machine learning (ML) and deep learning techniques for image recognition of various microorganisms. This review explores several key research inquiries, including image preprocessing, feature extraction methods, classification techniques, evaluation metrics, limitations of existing methods, and the evolution of technology over time [10]. Hosseininia, M. et al. introduced a novel approach for annotating 3D images using deep learning and view-based image features. One of the primary hurdles in automating the annotation of 3D images is the extraction of suitable features to represent these images. Unlike traditional 3D representations like polygonal meshes, which are less compatible with deep learning methods, this method leverages view-based features to enhance annotation accuracy [11].

In order to reduce the negative impact of current laser image pattern recognition methods on image applications, a laser image pattern recognition method based on big data analysis is proposed. Firstly, briefly explain the principles of big data analysis applied in the research process; Implement image analysis and training processing on the original laser image big data; Extract the features of laser images and perform pre segmentation on them to achieve pre-processing of the original laser images; Finally, the pattern recognition of laser images is completed using the regularized least squares method. Set up a simulation experiment to verify the application effect of this method. Using 10 types of laser images and a total of 1000 images as samples, set the preset target image recognition rate, recognition error rate, and image recognition consumption time as evaluation indicators for the simulation experiment. Apply this method for image pattern recognition analysis.

3. Method.

3.1. Overall Process of Design Methods.

The process of laser image pattern recognition based on big data analysis technology is shown in Figure 3.1.

According to the content in Figure 1, set the laser image pattern recognition process into three stages. In the first step, apply the foundation of big data analysis to train and process images, and use it as the basis for image pattern recognition [12,13]. Referring to computer vision technology and digital image processing technology, set up the laser image pattern recognition process to achieve the design goal. After the design of the identification method is completed, an experimental section is constructed to analyze the application effect of this method and clarify its advantages and disadvantages.

3.2. Principles of Big Data Analysis.

Compared with various big data analysis techniques, convolutional neural network technology was used to train and process the original laser images in this study [14]. Considering the unique attributes of laser images, formulas 3.1 and 3.2 have been derived to define the calculation framework of the convolutional neural network.

$$a(x,y) = h \times g = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(w,z)g(x-w,y-z)dwdz \tag{3.1}$$

$$A(i,j) = \sum_{m}\sum_{n} H(m,n)G(i-m,j-n) \tag{3.2}$$

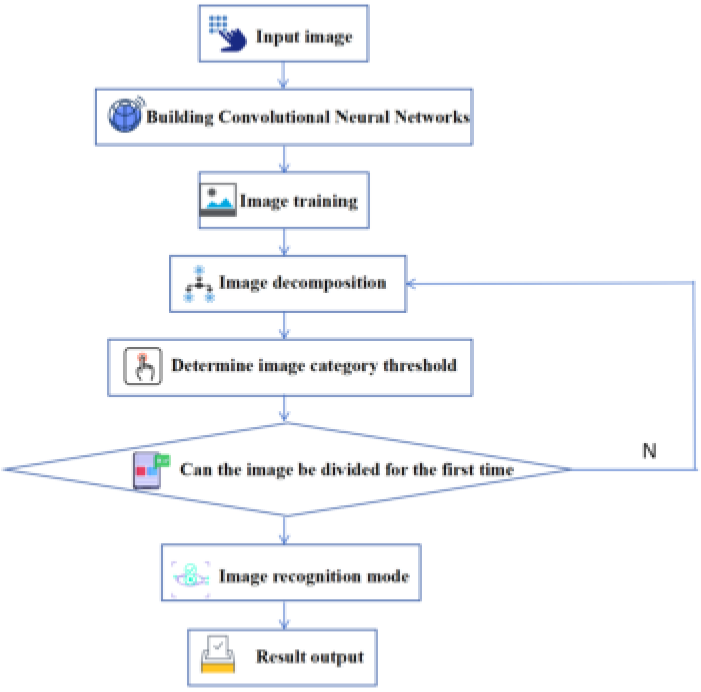

Input image
Building Convolutional Neural Networks
Image training
Image decomposition
Determine image category threshold
Can the image be divided for the first time
N
Image recognition mode
Result output

Fig. 3.1: Laser Image Pattern Recognition Process for Big Data Analysis

In equations 3.1 and 3.2, a(x,y) denotes the node position within the convolutional neural network. H signifies the convolutional calculation function employed. g represents the total number of layers in the convolutional neural network. h(w,z) corresponds to the original features of the image. d denotes the distance between nodes in the network. h(m,n) refers to the coordinate indices of the convolutional neural network nodes.

Using the above formula, perform convolution on the image to obtain the size of the processed convolution kernel. Then, based on this calculation result, obtain the length and width of the processed image. The specific value results are shown in equation 3.3:

$$\begin{cases} out_l = \frac{in_i}{stride_i} \\ out_w = \frac{in_w}{stride_w} \end{cases} \tag{3.3}$$

After determining the length and width of the image matrix, it is necessary to use this convolutional neural network again for secondary processing to remove noise from the image. In this process, the activation function of the neural network needs to be set, which can be expressed as equation 3.4:

$$f(a) = \frac{1}{1+\alpha^{-a}} \tag{3.4}$$

Control and optimize the calculation process of the neural network through this activation function for further processing.

3.3. Image Analysis Training Processing. After using convolutional neural networks for big data analysis of the original laser image, the analysis and training processing of the laser image begins [15]. In order to ensure the stability of the application effect of image convolutional neural networks, the constant during image training is set to $\eta = 10^{-8}$, the initial training parameter is set to λ, the first-order variable is initialized to e=0, the second-order variable is initialized to r=0, the initialization step is set to t=0, and the number of samples in the image training set is set to m'. At this time, the calculated gradient of the image training process can be expressed as equation 3.5:

$$s \leftarrow \frac{\nabla\sigma \sum_t K(f(x^i;\sigma), y^i)}{m} \tag{3.5}$$

After determining the gradient calculation, set the step size update formula as equation (6):

$$t \leftarrow t+1 \tag{3.6}$$

After determining the above calculation process, use the above content to train the image, in order to improve image quality and provide a foundation for subsequent image processing and recognition.

3.4. Laser image preprocessing. Based on the results of image training and processing, extract image features and set a threshold for pre segmentation. After comparing multiple methods, the OTSU threshold segmentation method was used to complete this part of the processing [16]. Set the pixel grayscale value range in laser image o (x, y) to , the probability of each grayscale value appearing to be set to , and the threshold to be set to T. This threshold is set to two categories according to the grayscale value range of pixel points, as shown in equation 3.7:

$$\begin{cases} L_0 = \{0, T\} \\ L_1 = \{T+1, Z-1\} \end{cases} \tag{3.7}$$

The probability of these two types of thresholds appearing can be expressed as equation 3.8:

$$\begin{cases} u_0 = \sum_{i=0}^{i} p' \\ u_1 = 1 - u_0 \end{cases} \tag{3.8}$$

According to formulas 3.7 and 3.8, the average grayscale of two types of laser images is obtained, expressed as equation 3.9:

$$\begin{cases} \zeta_0 = \sum_{i=1}^{i} \frac{ip'}{u_0,} = \frac{\zeta_i}{u_0} \\ \zeta_1 = \sum_{i=1}^{z-1} \frac{ip'}{u_0} = \frac{\zeta-\zeta_i}{1-u_0} \end{cases} \tag{3.9}$$

In equation 3.9,$\zeta = \sum_{i=1}^{z-1} ip'$,$\zeta_i = \sum_{i=1}^{t} ip'$.

After determining the threshold, mathematical morphology is used to determine the center of regions for different image types [17]. According to this threshold, the original image set is set to G, where H represents the structural elements of the image. Consider the process of image pre partitioning as setting a sliding window in the image and performing morphological calculations on each element in the graph. In this study, expansion and corrosion operations were mainly used for laser image features [18]. Expansion and corrosion operations are shown in equations 3.10 and 3.11 respectively:

$$G' \oplus H' = \{x|[\overline{H'}_x \cap G] = \varnothing\} \tag{3.10}$$

$$G' \oplus H' = \{x|\overline{H'}_x \subseteq G\} \tag{3.11}$$

Using the above operation, divide the image into two parts according to the preset threshold. After the image segmentation is completed, remove the images that cannot be partitioned, and store these two parts of the images in different databases [19].

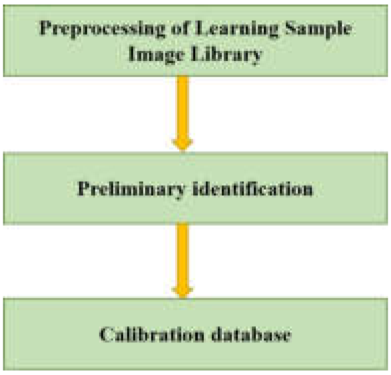

Preprocessing of Learning Sample Image Library
Preliminary identification
Calibration database

Fig. 3.2: Raw image data acquisition process

3.5. Laser Image Pattern Recognition. Use regular least squares method to analyze image elements in different databases and complete the laser image pattern recognition process [20]. In this study, the processed laser image pattern recognition process is considered as an image optimization problem, and the processing process can be set as equation 3.12:

$$\widetilde{z'} = argmin\frac{1}{u'}\sum_{i=1}^{u'} W'(y'_i - f(x'_i))^2 + J||z'||_i^2 \tag{3.12}$$

In equation 3.12, x'_i represents the feature vector of the image after initial segmentation; i represents the independent variable factor of the image sample; y_i represents the dependent variable data of the image sample.

The above calculations are all completed under the premise that the image sample set contains u' samples, where $||z'||_i^2$ represents the norm of the function z' induced by the calculation process. According to this calculation principle, on the premise of determining the image target pattern, the pattern recognition process is integrated into the form shown in equation 3.13:

$$max = v_0(\gamma_0 - \gamma_i)^2 + v_1(\gamma_1 - \gamma_i)^2 \tag{3.13}$$

In equation 3.13, the expansion forms of $v_0, v_i, \gamma_0, \gamma_1$, and γ_i are shown in equation 3.14:

$$\begin{cases} v_0 = \sum_{i=0}^{t-1} b'(i), v_o = \sum_{i=0}^{t-1} b'(i) \\ \gamma_0 = \frac{\sum_{i=0}^{t-1} ib'(i)}{v_0}, \gamma_1 = \frac{\sum_{i=0}^{t-1} ib'(i)}{v_1}, \gamma_i = \sum_{i=0}^{t-1} ib'(i) \end{cases} \tag{3.14}$$

In equation 3.14, represents the target mode image features; v_1 represents non target mode image features; γ_0 represents the information value of the target image library after the initial segmentation; γ_1 represents the information value of the non target image library after the initial segmentation; γ_i represents the information value of the calculation result obtained from formula 3.12; b' represents the image pattern segmentation coefficient.

After the calculation of formula 3.13 is completed, the final division is performed using the binary method, and the target mode can be expressed as equation 3.15:

$$E' = E_0 \times b' \tag{3.15}$$

The process of collecting raw data is shown in Figure 3.2.

Organize the above calculation steps to obtain the final laser image pattern recognition result.

By integrating the content set in the previous text, the laser image pattern recognition method based on big data analysis has been set up.

Table 3.1: Sample Division Results of Laser Image Experiment

Experimental Image Set Number	Number of images	Main image type
CY-01	100	X-10
CY-02	100	X-3
CY-03	100	X-4
CY-04	100	X-6
CY-05	100	X-7
CY-06	100	X-8
CY-07	100	X-9
CY-08	100	X-5
CY-09	100	X-1
CY-10	100	X-2

3.6. Experimental Analysis. In this study, 10 types of laser images were used as sample images, which can be divided into 4 categories, with a total of 1000 images. The laser image type serial numbers are set to X-1 to X-10, corresponding to 10 different image types. Due to the different sources of experimental images, there are certain differences in the accuracy of the images. To avoid this difference affecting the image recognition results, the images are converted into high-dimensional image information and input into the computer to complete the recognition process.

In order to improve the contrast of the experiment, the original images were summarized and divided into 10 experimental image groups. Each experimental group also had 50 corresponding laser images of its own group and 5 different types of laser images, with 10 laser images in each group. The specific experimental group division results are shown in Table 3.1.

Organize the above settings and import them into the experimental platform to provide a foundation for the subsequent experimental process.

The design of this method is mainly aimed at strengthening the training process of the original image in laser image pattern recognition, improving the ability of image pattern recognition, and thereby improving the recognition accuracy for different types of laser images. Therefore, based on previous experimental research results, the evaluation indicators of image pattern recognition methods are set as preset target image recognition rate, recognition error rate, and recognition accuracy in the experimental section. They can be specifically expanded as follows:

(1). Default target image recognition rate:

$$T' = \frac{R_1}{R_{all}} \times 100\% \tag{3.16}$$

In equation 3.16, R_1 represents the number of recognized target images; R_{all} represents the number of target type images set.

(2). Recognition error rate

$$D = \frac{(N - N_{all})}{N_{all}} \times 100\% \tag{3.17}$$

In equation 3.17, N represents the number of correctly recognized pattern images; Nall represents the total number of various pattern images present in the experimental image set.

By converting the selected laser image into high-dimensional image information and inputting it into the computer, the recognition process can be completed. The start and end recognition times can be obtained, and the difference between the two can be calculated to obtain the recognition consumption time for each experimental image. Further select image recognition methods such as enhanced canonical correlation analysis, dynamic near-infrared spectroscopy, and feature vector extraction to compare with this method, and analyze their advantages and disadvantages.

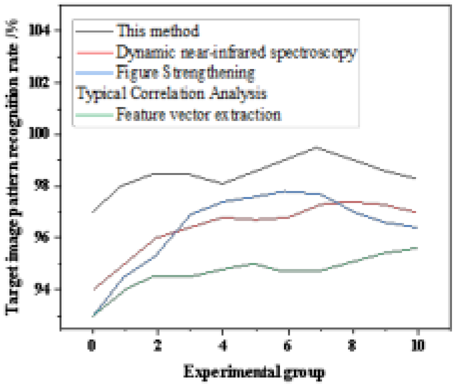


Fig. 4.1: Experimental results of preset target image recognition rate

4. Results and Discussion. The experimental results of the preset target image recognition rate are shown in Figure 4.1. Upon examining the findings presented in Figure 3, it is evident that this method achieves a notably high recognition rate for the predefined target images. Implementing this approach can effectively ensure satisfactory image recognition outcomes. Compared with this method, the image recognition performance of feature vector extraction is relatively poor, with a recognition rate of less than 95.00%. This method cannot obtain high-quality image recognition results. The target image recognition rates of the other two methods are higher than those of neural networks, but the overall volatility is high and the usage effect is unstable. Based on the above experimental results, it can be preliminarily determined that the application effect of this method should be superior to the other three methods.

The experimental results of recognition error rate are shown in Figure 4.2. In this study, this indicator was used to verify the recognition accuracy of image pattern recognition methods for different pattern images. After the experiment is completed, plot the experimental data as shown in Figure 4. From the analysis of the above images, it can be seen that this method has a relatively low recognition error rate for different pattern images, indicating that this method can be used to finely divide images and avoid unclear classification of image categories. Compared with this method, the error rates of different pattern image recognition for the other three methods are significantly higher than this method. This experimental result confirms that big data analysis technology has a certain recognition accuracy in this method, and the application of this technology can improve the discrimination effect of different types of information to a certain extent.

Figure 4.3 shows the experimental results of image recognition time consumption. Analysis of the image content in Figure 5 shows that there are significant differences in the recognition time of the sample images among the four methods. The recognition time of this method is significantly shorter than the other three methods, and the overall time is smaller and tends to be stable. The other three methods have higher recognition time for some groups, but the recognition time for some groups is not ideal. By organizing the above content and combining the experimental results of preset target image recognition rate and recognition error rate, it can be determined that this method is the best image recognition method among the experimental methods.

This method can ensure a preset target image recognition rate of over 96% for different types of laser images, which is significantly better than the three comparison methods; The overall stability of the recognition error rate is below 2%. Indicating that the application effect of this method is good and the accuracy of image recognition is high. Finally, the time consumption for pattern recognition of different types of laser images under various methods was counted, and it was found that the average time consumption of this method was 9.4 seconds, which belongs to a lower level, confirming that this method has a shorter recognition time and higher efficiency. With the continuous expansion of the application scope of laser images, this method provides impetus for the improvement of laser image analysis and processing technology.

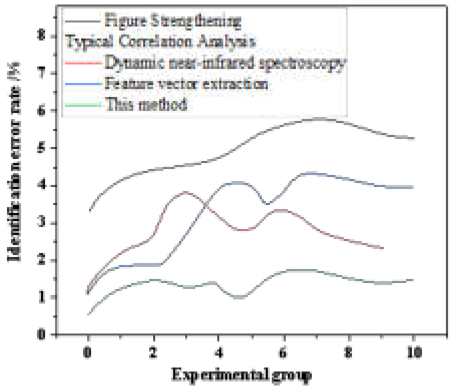


Fig. 4.2: Experimental results of recognition error rate

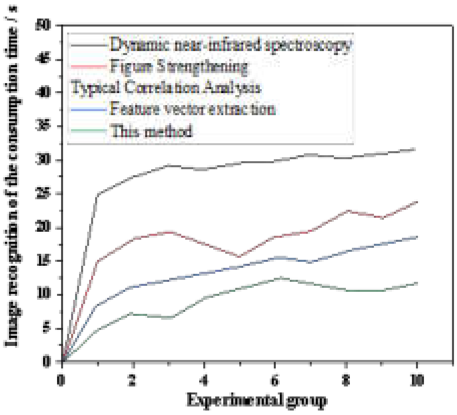


Fig. 4.3: Experimental results of image recognition time consumption

5. Conclusion. The author proposes the application research of deep learning based image recognition technology in data analysis. To minimize recognition errors in laser image recognition methods and enhance accuracy in this domain, the research proposes a method leveraging big data analysis technology. Utilizing convolutional neural networks ensures stability in learning from ample data, thereby improving recognition accuracy. Training and processing the raw laser image big data can effectively enhance its performance in the image training stage, and the above steps have laid a solid foundation for improving the accuracy of laser image pattern recognition.

Acknowledgments. The work was supported by 2025 Henan Province Science and Technology Department Key Research and Promotion Special Project (Science and Technology Research) Research on path planning for autonomous valet parking based on reinforcement learning (No.252102210226).

REFERENCES

[1] Wang, T. (2022). Exploring intelligent image recognition technology of football robot using omnidirectional vision of internet of things. The Journal of Supercomputing, 78(8), 10501-10520.

[2] Shen, H., Huo, K., & Qiao, X. L. C. (2023). Aircraft target type recognition technology based on deep learning and structure feature matching. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 45(4), 5685-5696.

[3] Wang, Y., Yang, Z., Li, Z., Tian, X., Zhai, L., & Gu, S., et al. (2022). A novel fingerprint recognition method based on a siamese neural network. Journal of Intelligent Systems, 31(1), 690-705.

[4] He, L., Guo, C., Su, R., Tiwari, P., Pandey, H. M., & Dang, W. (2022). Depnet: an automated industrial intelligent system using deep learning for video-based depression analysis. International Journal of Intelligent Systems, 37(7), 3815-3835.

[5] Nagaeva, D. Z., Sazonova, E. Y., Smetanina, O. N., & Sazonov, V. S. (2023). Technology of knowledge engineering in the diagnosis of eating behavior. Pattern Recognition and Image Analysis, 33(3), 452-459.

[6] Qingjian, L. I., Yan, L., Zhenzhou, L. U., & Guangyi, W. (2023). Threshold-type memristor-based crossbar array design and its application in handwritten digit recognition, 34(2), 324-334.

[7] Leiloglou, M., Kedrzycki, M. S., Chalau, V., Chiarini, N., Thiruchelvam, P. T. R., & Hadjiminas, D. J., et al. (2022). Indocyanine green fluorescence image processing techniques for breast cancer macroscopic demarcation. Scientific reports, 12(1), 8607.

[8] Maiti C., Muthuswamy S. (2024). Classification of materials in cylindrical workpieces using image processing and machine learning techniques. International Journal of Production Research, 62(7), 2566-2583.

[9] Kovalerchuk, B., Kalla, D. C., & Agarwal, B. (2022). Deep learning image recognition for non-images, 34(1), 249-270.

[10] Rani, P., Kotwal, S., Manhas, J., Sharma, V., & Sharma, S. (2022). Machine learning and deep learning based computational approaches in automatic microorganisms image recognition: methodologies, challenges, and developments. Archives of computational methods in engineering: State of the art reviews(3), 29.

[11] Hosseinnia, M., & Behrad, A. (2023). 3d image annotation using deep learning and view-based image features. 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA), 1-6.

[12] Sadigh, S., & Kim, A. S. (2024). Molecular pathology of myeloid neoplasms: molecular pattern recognition. Clinics in Laboratory Medicine, 44(2), 339-353.

[13] Tao, X. U., Huan, Y. U., Xia, Q., Bo, K., Qing, X., & Xiaoyu, X. U., et al. (2023). Analysis of morphological characteristics of gravels based on digital image processing technology and self-organizing map, 15(3), 17.

[14] Park, J. I., Koo, C. U., Oh, J., Kim, I. J., Choi, K., & Ye, S. J. (2023). Enhancing precision in l-band electron paramagnetic resonance tooth dosimetry: incorporating digital image processing and radiation therapy plans for geometric correction. Health Physics, 126(2), 79-95.

[15] Bal-Onazi, B., Alotaibi, N., Salzahrani, J., Alshahrani, H., Elfaki, M. A., & Marzouk, R., et al. (2023). Modified dragonfly optimization with machine learning based arabic text recognition, 76(8), 1537-1554.

[16] Liu, S., Wu, J., Zhang, X., Zhao, H., Li, X., & Hu, R., et al. (2022). Research on data classification and feature fusion method of cancer nuclei image based on deep learning. International Journal of Imaging Systems and Technology, 32(3), 969-981.

[17] Lu, F., Xu, L., Jiang, Y., Luo, P., Hu, X., & Liang, J., et al. (2022). Smear character recognition method of side-end power meter based on pca image enhancement. Nonlinear Engineering, 11(1), 232-240.

[18] Fan, R., & Han, X. (2024). Deep palmprint recognition algorithm based on self-supervised learning and uncertainty loss. Signal, Image and Video Processing, 18(5), 4661-4673.

[19] Althabhawee, A. F. Y., & Alwawi, B. K. O. C. (2022). Fingerprint recognition based on collected images using deep learning technology. IAES International Journal of Artificial Intelligence, 11(1), 81-88.

[20] Hyochang, A., & Han-Jin, C. (2023). Research of automatic recognition of car license plates based on deep learning for convergence traffic control system. Personal and ubiquitous computing, 27(3), 1139-1148.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 1, 2024
Accepted: Jul 31, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1466–1473, DOI 10.12694/scpe.v26i3.4269

DESIGN OF VIRTUAL ROAMING SYSTEM OF ART MUSEUM BASED ON VR TECHNOLOGY

JIA YANG[*] AND XIAYING WU[†]

Abstract. This paper aims to explore a panoramic Mosaic algorithm combining particle swarm optimization (PSO) and mutual information (MI) to improve the immersion and interactive performance of the virtual tour system of art museums. First, this paper introduces the application background of VR technology in the virtual art museum tour and emphasizes its importance in breaking the limitation of time and space and enhancing the audience's participation. Then, the application of the particle swarm optimization algorithm in image registration is described in detail. By simulating the foraging behavior of birds, the algorithm effectively solves the matching problem in the process of panoramic Mosaic. At the same time, mutual information is introduced as an index to evaluate image similarity, which further improves the accuracy and efficiency of stitching. Then, this paper proposes a virtual roaming framework based on the panoramic Mosaic algorithm, which can seamlessly integrate high-resolution artwork images and realize free navigation in 3D space through VR headsets. In addition, the system also supports various interaction modes, such as gesture control, speech recognition, etc., to meet the needs of different users. Finally, through a model simulation test, this paper shows the significant advantages of the designed virtual roaming system regarding visual effects and user experience. The experimental results show that the system can provide a highly realistic exhibition environment and enhance the audience's immersion through intelligent interaction. The system provides new ideas and technical support for the digital transformation of art museums.

Key words: VR technology, Art Museum virtual tour system, Particle swarm optimization algorithm, Mutual information, Panorama mosaic algorithm, Model simulation

1. Introduction.

Under the tide of the digital age, virtual reality (VR) technology is gradually penetrating the temple of culture and art - art museums with its unique immersive experience and interactivity. The development of VR technology has brought revolutionary changes to the display methods of art museums, enabling the audience to enjoy the art feast across time and space without being restricted by physical space.

In recent years, the application of VR technology in the virtual tour system of art museums has become a research hotspot. Literature [1] puts forward the concept of creating a virtual art museum using VR technology, which solves the problem that traditional art museum visits are limited by geographical location and opening hours. However, improving the realism and interactivity of virtual roaming has become an urgent problem for researchers. The particle swarm optimization algorithm (PSO) was introduced in the literature [2], which optimized the image registration process by simulating the swarm intelligence behavior, thus improving the quality of the panoramic Mosaic. However, the PSO algorithm still has some limitations when dealing with complex scenes. Literature [3] proposed a method combining mutual information (MI) to evaluate the similarity by calculating the mutual information between images, further improving image stitching accuracy. However, how to achieve an efficient panoramic Mosaic while ensuring image quality is still the focus of current research.

In addition, the selection of the panorama Mosaic algorithm directly impacts the performance of the virtual roaming system. Literature [4] compared several mainstream panorama Mosaic algorithms and pointed out that the algorithm based on feature point matching has advantages in speed and accuracy. However, these algorithms often face the problem of large consumption of computing resources when processing large-scale image data. Therefore, literature [5] proposed a GPU-based panorama Mosaic algorithm, which effectively improved the processing speed, but the compatibility and stability of this method on mobile devices still need to be improved.

[*] Guangdong Vocational Institute of Public Administration, Guangzhou 510800, China (Corresponding author, hajimeyoung@163.com)

[†] Guangdong Agriculture Industry Business Polytechnic, Guangzhou 511365, China

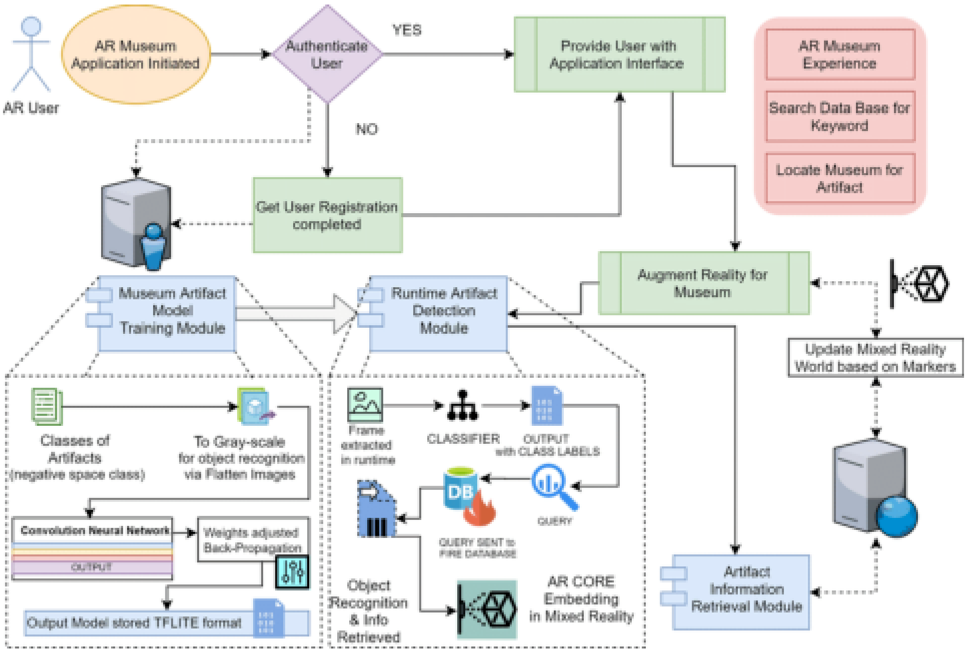

AR User
AR Museum Application Initiated
Authenticate User
YES
NO
Provide User with Application Interface
AR Museum Experience
Search Data Base for Keyword
Locate Museum for Artifact
Get User Registration completed
Augment Reality for Museum
Museum Artifact Model Training Module
Runtime Artifact Detection Module
Update Mixed Reality World based on Markers
Classes of Artifacts (negative space class)
To Gray-scale for object recognition via Flatten Images
Convolution Neural Network
OUTPUT
Weights adjusted Back-Propagation
Output Model stored TFLITE format
Frame extracted in runtime
CLASSIFIER
OUTPUT with CLASS LABELS
DB
QUERY
QUERY SENT to FIRE DATABASE
Object Recognition & Info Retrieved
AR CORE Embedding in Mixed Reality
Artifact Information Retrieval Module

Fig. 2.1: Architecture diagram of virtual Art Museum system.

This paper will synthesize the research results of the above literature to design a new type of museum virtual roaming system [6]. Firstly, this paper will use particle swarm optimization algorithm and mutual information to optimize the panoramic Mosaic process to achieve higher image quality and faster processing speed. Secondly, this paper will explore a splicing algorithm based on GPU acceleration to adapt to the performance requirements of different devices. This paper will then examine how to enhance user immersion and engagement through intelligent interaction technologies such as gesture recognition and voice control [7]. Finally, this paper will evaluate the performance of the designed system through model simulation and user experiments and put forward suggestions for improvement.

2. System architecture design of virtual art museum.

2.1. System Architecture. The virtual art museum contains a three-layer structure: data storage, core business, and display layers (Figure 2.1 is quoted in Using Augmented Reality and Deep Learning to Enhance Taxila Museum experience). The data storage layer manages and stores system model data, attribute data and spatial data in a unified manner to achieve the purpose of data integration management. The system includes the main commercial functions, such as map service, permission management, information query, 3D model display, map annotation, browser adjustment, etc. Some modules can be built separately with software such as 3DS MAX. The function of the presentation layer is to display the user's interactive interface and 3D virtual environment online. Using VRML technology, the 3D virtual environment model can be displayed in the network. VRML is also inserted into a web page to view a 3D virtual environment.

2.2. System function design of virtual art museum. The function of the virtual art museum is to present the art museum to the public in the form of pictures through the introduction of the art museum, the operation of the two-dimensional map of the art museum, the query of works information, and the visit of the virtual three-dimensional art museum [8]. This is of great help to enhance the visibility of the museum. The specific functional structure of the system is divided into the following parts: (1) Art Museum introduction. Its content is a simple description of the general situation of the gallery through the way of text, including the origin of the gallery, the purpose of the museum, the collection of information, business hours, etc., so that more people have a better understanding of the gallery and even the museum. (2) Mapping operation. The content of map operation is the essential operation of a geographic information system, which includes processing 2D plan enlargement, reduction, translation, selection map annotation, etc. Its role is to display the layout of an art museum and the layout of works. (3) Information inquiry. The "information query" function focuses on retrieving paintings, while in the graphic gallery design, the paintings are represented by a dot or a rectangular frame. When you use the Information Query function, a window to draw the image appears. (4)3D roaming. The most important feature of this system is the "three-dimensional browsing" function. In the "3D Browse", a page will appear, reminding users to download the VRML plug-in and then complete the 3D tour of the museum on this page. This gives the user a feeling of being there. (5) Authority management. The "Rights management" module ensures the system's security. You can manage groups of users that have logged in to the system and assign login and management rights to different groups. (6) Tourist information bar. People can visit the gallery after the experience and target to give their views and suggestions so that other visitors in the tour process to play a specific role in the reference, but also can give museum managers some service suggestions to make the museum service more perfect.

2.3. Specific modeling of virtual art museum. The virtual gallery is divided into 5 modules: wall model, top component model, top lighting model, ceiling model and individual painting model. After accurately measuring the walls and door frames of the museum, different rectangular models are made according to the predetermined space dimensions [9]. The top and wall elements constitute the overall shape of the entire gallery perimeter, and the coding is roughly the same as that of the wall. A series of miniature lighting fixtures are made up of tin cylinders and spherical balls arranged on the ceiling. The ceiling model consists of the top and the floor where the lamps are arranged. The Indexed Face constructs the top, and the floor where the lamps are laid is extended by an Extrusion pattern. The pre-made lighting model is regularly displayed on the floor in fixed coordinates, and the spacing of the coordinates is balanced according to the spacing of the X, Y, and Z axes. Keep them as evenly spaced as possible. The individual painting models are the main exhibits in a virtual gallery, and they need to be photographed and imported into a pre-made rectangular frame and drawing board. Using glass materials to make the paintings appear.

2.4. Model data storage and presentation. First, each module in the completed art gallery is formed into an independent solid model. The space layout scheme of the 3D gallery is designed, and the saved 3D stereo models are added to the 3D space one by one and integrated with VRML documents [10]. This results in a complete three-dimensional art museum (Figure 2.2). After the VRML document is generated, the various virtual environments are integrated using Inline nodes to speed up web browsing and download rates.

3. Use mutual information combined with the PSO algorithm to achieve a panoramic image Mosaic.

3.1. Mosaic of panoramic images. The image is preprocessed by a smoothing method to reduce the interference of noise [11]. The PSO algorithm is used to find the features in as many images as possible, and the image matching is found by using the mutual information degree to complete the Mosaic. Here are the specific steps:

3.1.1. Image smoothing. Some random noise is often generated during imaging because the appearance of these noises will cause the details of the adjacent image to be not wholly consistent, which will have a more significant impact on the later processing. Therefore, this article should smooth the image first. The image is smoothed according to the following equation.

$$P(i,j)=\sum_{m=0}^{T-1}\sum_{n=0}^{N-1}L(m,n)W(i-m,j-n) \tag{3.1}$$

W is the input image. P represents the output image. L is a smooth convolution filtering algorithm. Gaussian smooth filtering is used here.

3.1.2. PSO Algorithm. The traditional image registration is mainly used to find the exact overlap between two images. In general, search for A template $\mathrm{lm}\, f_T$ in picture N. Use this template to match the image S. Finally, the two images were matched to achieve registration (Figure 3.1 cited in Comparison of

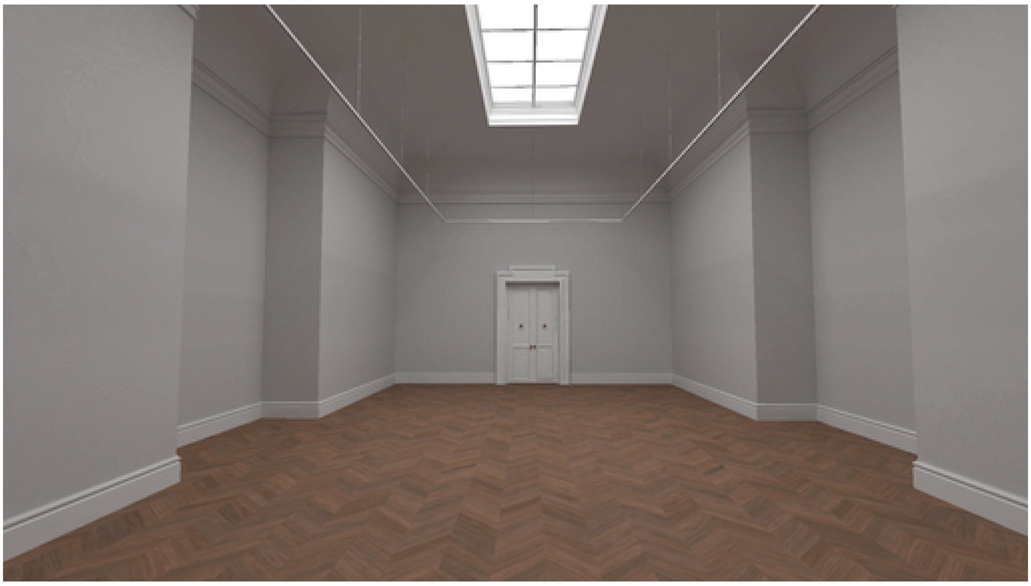


Fig. 2.2: Example of the final integration model of the museum.

Population Based Intelligent Techniques to Solve Load Dispatch Problem). Search for $\operatorname{lm} f_T$ using PSO. The most significant advantage of PSO is its rapid convergence. A multi-target detection method based on wavelet transform is proposed in this paper. Match other images S to achieve alignment. First, the particles are initialized. Both positioning and speed are included [12]. Then, the fitness of each particle is calculated. The method of iterative optimization is adopted. Each particle moves on its own. One is the personal limit q_{best}. The other is the overall maximum f_{best} . The performance ends when sufficient good fit or maximum number of iterations is reached. The particle self-renews its rate and orientation according to the following formula:

$$\begin{aligned} & U(\tau+1)=U(\tau)+\operatorname{rand}() \times z_1 \times\left(q_{b \text { bes }}(\tau)-\operatorname{present}(\tau)\right) \\ & +\operatorname{rand}() \times z_2 \times\left(f_{\text {best }}(\tau)-\operatorname{present}(\tau)\right)\end{aligned} \tag{3.2}$$

$$\operatorname{present}(\tau+1)=\operatorname{present}(\tau)+U(\tau+1) \tag{3.3}$$

Here $U(\tau)$ is the particle rate of time $\tau \cdot \operatorname{present}(\tau)$ is the location of the current particle on time b. rand() is any number chosen in E. F is the learning coefficient, and in general, $z_1 = z_2 = 2$. Ten particles are randomly allocated in the first picture's right half of the space using the 50% overlap principle [13]. The coordinates of several pixel points determine an initial value. Set the initial particle speed.

3.2. Mutual Information. If people continue to use the conventional N_2 distance, the calculation is complicated and easy to be interfered with by lighting, brightness, and other factors, but also, when there is a specific Angle deviation between the two frames, the algorithm will often fail. Because of the robustness of the mutual information measure, this paper uses it as a statistical correlation measure between images [14]. The mutual information of the algorithm is like that of the algorithm without rotation error when there is a slight rotation deviation between two images. The algorithm is suitable for statistical correlation analysis of two images in various cases. Cross-information is a measure used to describe the relationship between random variables [15]. Assume that the image N and S are two uncertain variables. $L(N)$ is the entropy of picture N, and $L(S)$ is the entropy of picture S. Where $L(N, S)$ is their total entropy. So, the amount of mutual information between them is:

$$Y(N, S)=L(N)+L(S)-L(N, S) \tag{3.4}$$

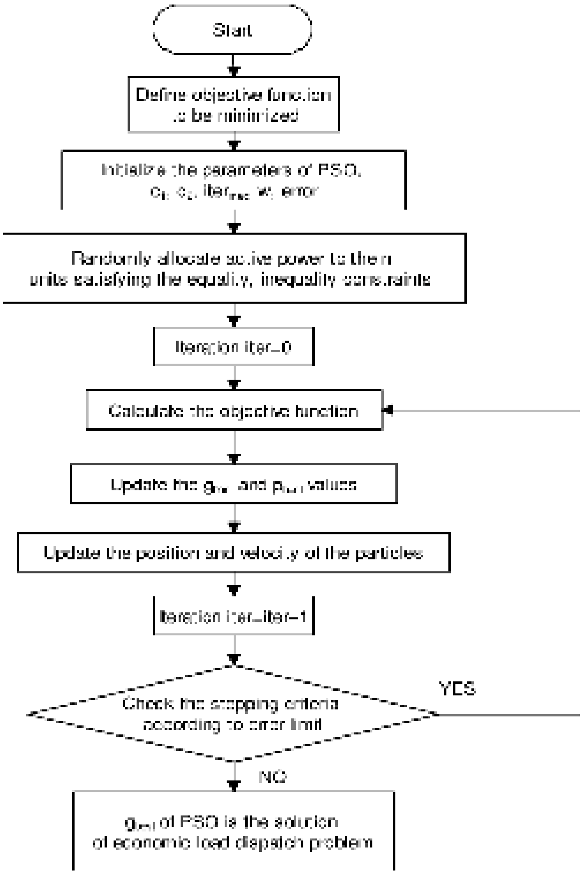

Start
Define objective function to be minimized
Initialize the parameters of PSO, c_1, c_2, $iter_{max}$, w, error
Randomly allocate active power to the n units satisfying the equality, inequality constraints
Iteration iter=0
Calculate the objective function
Update the g_{best} and p_{best} values
Update the position and velocity of the particles
Iteration iter=iter+1
Check the stopping criteria according to error limit
YES
NO
g_{best} of PSO is the solution of economic load dispatch problem

Fig. 3.1: PSO algorithm flow.

If the gray probability density distributions of images N and S are $q_N(l)$ and $q_S(s)$, respectively. The gray joint probability density distribution is $P_{NS}(l,s)$. Mutual information $Y(N,S)$ can be expressed as:

$$Y(N,S)=\sum_{a,b} q_{NS}(l,s)\log\frac{q_{NS}(l,s)}{q_N(l)q_S(s)} \tag{3.5}$$

Entropy is:

$$L(X)=\sum_{x} q_X(x)\log q_X(x) \tag{3.6}$$

The joint entropy is:

$$L(X,Y)=-\sum_{x,y} q_{MY}(x,y)\log q_{XY}(x,y) \tag{3.7}$$

In the left half of image S, find a zone $\text{Local}\,\text{lm}\, f_i i=1,2,\cdots,n$ of the same size as the template $\text{lm}\, f_T$. Compare the mutual information between this region and lm f_T template. The most mutually informative region found in image S is used as a matching region $\text{lm}\, f_T'$. Then, image S and image N can be registered according to the matching region. Thus, registration based on mutual information metrics can be expressed as:

$$w'=\arg\max_{w} Y\left(\text{lm}\, f_T, w\left(\text{ Local }\ \text{lm}\, f_i\right)\right) \quad i=1,2,\cdots,n \tag{3.8}$$

4. System implementation.

4.1. Model standardization processing.

4.1.1. UVW and stickers. UVW uses modeling software to express the three-dimensional space of the X, Y and Z axes. For texture, U is the component in the horizontal direction, and V is the component in the vertical direction. Before adding materials and textures, they must be UVW adjusted so that subsequent materials or textures can achieve the desired results on each component. Texture mapping can be divided into image and programming textures [16]. The image texture is adding the processed image to the model's surface. The algorithm is relatively simple. However, if it is enlarged, there are apparent arrangement tracks and partial Mosaic phenomenon. Therefore, for the accuracy and repeatability of images, maps are generally only suitable for regular brick or independent flat textures. Program mapping is a calculation method by the software, according to the set parameters to produce a material, because of its vector and disorder, so whether it is scaled or shrunk, there will be no Mosaic and no gaps [17]. But, because its structure is cumbersome and limited, its application scope is insignificant. As for the texture method used, it depends on the actual needs and the requirements for accuracy.

4.1.2. Model Optimization. Because the algorithm needs to complete the tasks of illumination information statistics and shadow calculation in the calculation process, when the number of models is too large, the memory cost will increase in the calculation process, resulting in screen delay, response speed and other problems. Therefore, polygons in the following cases can be deleted: (1) the intersection area of the grid is divided; (2) Polygons that will not exist in the image; (3) Over-detailed polygons and so on.

4.1.3. Model Export. After the modeling is completed, the model of the same material is spliced according to the commonness and difference of the material, Adding automatic smoothing features to the surface layers of all models. A painting can combine parts and remove redundant layers or patterns from the solution manager. Name a single Grid body, image, material, etc., generally M_ (material name), T_ (map name), SM_ (static Grid body name), etc. After all the above operations are done, select all the modes, select the coordinates in the right toolbar, and specify the XYZ coordinates of the mode to 0. Select Export Selection Mode from the file options in the upper left corner. Export the pattern to the working folder [18]. Select Smooth Group above the output TAB and cancel Camera, Animation, and Lights.

4.2. Available Functions. Using Steam VR and VRTK technology, people completed the Teleportation, slow walking, scene interactive jumping and display of work information in 3 scenes.

4.2.1. Teleportation. Through virtual reality technology, the player can move to the designated position instantaneously by manipulating the light on the stick. After introducing the Steam VR plug-in and VRTK plug-in into the Unity3D engine, create a blank target, rename it VRTK_Manager, load the VRTK_SDK_Manager script, and add a blank object. VRTK_Controlller Events are added, and new control point references are defined separately in the Script Aliases column of VRTK_SDK Manager, thus avoiding damage to the original Camera Rig prefab. Under Camera Rig, select a control point to use as a teleportation indicator and add VRTK_ control events to this control point [19]. This code is used to listen for input from the HTC Vive controller. Then, add the VRTK_Pointer script to the control point and check the remote selection box.

In the destination setting event, the teleport flag bit is valid, so the transport script can decide whether to move to the new destination [20]. If this option is not selected, the lever emits a beam of light but does not trigger displacement. When the object tag is set to VRTK_Policy List, it cannot be transferred to the object.

4.2.2. The jump of the world in the painting. In the main stadium, there is a jumping area. Visitors step into the area, will be activated, and then teleported to a specific screen. After visiting the ”World in Picture,” press the ”back” button to return to the game site. This function is done by jumping around C# scripts.

4.2.3. Display of work information. When you visit the main museum, you can touch the light of your hand to the oil painting on the wall, and information about a work will appear [21]. Add a Canvas widget under each image, load the VRTK_UI Canvas script, and adjust the button size under the Canvas to cover the entire screen.

4.2.4. Scenario Optimization. Using particle effects, constructing terrain, and other means to render landscape and artistic atmospheres. Since the gallery is an indoor environment, its effect is achieved by lighting some details around the artwork and lighting from the top of the gymnasium. For roof lighting, add a Spot Light to the scene, adjust the brightness and direction of the lighting, and let the light fall on the place or artwork to be lit. For the detailed lighting in the works of art, this paper can combine the point light source so that every Angle of the work can get light exposure. It also prevents the light from getting too dark. In addition, to save calculation, the mode of part of the Light source can be converted to Baked, and then the Light Mapping technology is used to ”bake” the color and intensity information of the light source into the light mapping.

5. Conclusion. First, VR technology provides an unprecedented immersive experience for the museum's virtual roaming system, enabling the audience to cross geographical boundaries and enjoy the art journey anytime and anywhere. By integrating particle swarm optimization (PSO) and mutual information (MI) algorithms, this paper significantly improves the accuracy and efficiency of image processing and lays a solid foundation for constructing a virtual environment. Secondly, the system design scheme proposed in this paper shows excellent performance in the model simulation, which is close to the real art museum in terms of visual effect and realizes a qualitative leap in interactive experience. By integrating intelligent interactive technologies, such as gesture control and voice recognition, users can interact with virtual artworks more naturally, enhancing the fun and educational value of visits. However, despite these achievements, this study also reveals the challenges of existing technologies in handling complex scenes, improving rendering speed, and optimizing user experience. Future research should focus on further improving the robustness and adaptability of the algorithm and exploring more efficient image processing and rendering techniques to meet the growing visual and interactive needs.

REFERENCES

[1] Resta, G., Dicuonzo, F., Karacan, E., & Pastore, D. (2021). The impact of virtual tours on museum exhibitions after the onset of covid-19 restrictions: visitor engagement and long-term perspectives. SCIRES-IT-SCIentific RESearch and Information Technology, 11(1), 151-166.

[2] Mu, M., Dohan, M., Goodyear, A., Hill, G., Johns, C., & Mauthe, A. (2024). User attention and behaviour in virtual reality art encounter. Multimedia Tools and Applications, 83(15), 46595-46624.

[3] Othman, M. K., Nogoibaeva, A., Leong, L. S., & Barawi, M. H. (2022). Usability evaluation of a virtual reality smartphone app for a living museum. Universal Access in the Information Society, 21(4), 995-1012.

[4] Durmuş, U., & Günaydın, M. (2024). Virtual Reality Based Decision Support Model for Production Process of Museum Exhibition Projects. International Journal of Human–Computer Interaction, 40(11), 2887-2904.

[5] El-Said, O., & Aziz, H. (2022). Virtual tours a means to an end: An analysis of virtual tours' role in tourism recovery post COVID-19. Journal of Travel Research, 61(3), 528-548.

[6] Trunfio, M., Lucia, M. D., Campana, S., & Magnelli, A. (2022). Innovating the cultural heritage museum service model through virtual reality and augmented reality: The effects on the overall visitor experience and satisfaction. Journal of Heritage Tourism, 17(1), 1-19.

[7] Giannini, T., & Bowen, J. P. (2022). Museums and Digital Culture: From reality to digitality in the age of COVID-19. Heritage, 5(1), 192-214.

[8] Kim, Y., & Lee, H. (2022). Falling in love with virtual reality art: A new perspective on 3D immersive virtual reality for future sustaining art consumption. International Journal of Human–Computer Interaction, 38(4), 371-382.

[9] González-Rodríguez, M. R., Díaz-Fernández, M. C., & Pino-Mejías, M. Á. (2020). The impact of virtual reality technology on tourists' experience: a textual data analysis. Soft Computing, 24(18), 13879-13892.

[10] Casadio, F. (2021). Sharing power: Leadership lessons from interdisciplinary practices in an art museum. Curator: The Museum Journal, 64(3), 505-527.

[11] Raimo, N., De Turi, I., Ricciardelli, A., & Vitolla, F. (2022). Digitalization in the cultural industry: evidence from Italian museums. International Journal of Entrepreneurial Behavior & Research, 28(8), 1962-1974.

[12] Go, H., & Kang, M. (2023). Metaverse tourism for sustainable tourism development: Tourism agenda 2030. Tourism Review, 78(2), 381-394.

[13] Wu, W. L., Hsu, Y., Yang, Q. F., Chen, J. J., & Jong, M. S. Y. (2023). Effects of the self-regulated strategy within the context of spherical video-based virtual reality on students' learning performances in an art history class. Interactive Learning Environments, 31(4), 2244-2267.

[14] Trunfio, M., Campana, S., & Magnelli, A. (2020). Measuring the impact of functional and experiential mixed reality elements on a museum visit. Current Issues in Tourism, 23(16), 1990-2008.

[15] Guo, K., Fan, A., Lehto, X., & Day, J. (2023). Immersive digital tourism: the role of multisensory cues in digital museum experiences. Journal of Hospitality & Tourism Research, 47(6), 1017-1039.

[16] Wu, X., Chen, X., Zhao, J., & Xie, Y. (2024). Influences of design and knowledge type of interactive virtual museums on learning outcomes: An eye-tracking evidence-based study. Education and Information Technologies, 29(6), 7223-7258.

[17] Agostino, D., Arnaboldi, M., & Lampis, A. (2020). Italian state museums during the COVID-19 crisis: from onsite closure to online openness. Museum Management and Curatorship, 35(4), 362-372.

[18] Spachos, P., & Plataniotis, K. N. (2020). BLE beacons for indoor positioning at an interactive IoT-based smart museum. IEEE Systems Journal, 14(3), 3483-3493.

[19] Garbutt, M., East, S., Spehar, B., Estrada-Gonzalez, V., Carson-Ewart, B., & Touma, J. (2020). The embodied gaze: Exploring applications for mobile eye tracking in the art museum. Visitor Studies, 23(1), 82-100.

[20] Walsh, D., Hall, M. M., Clough, P., & Foster, J. (2020). Characterising online museum users: a study of the National Museums Liverpool museum website. International Journal on Digital Libraries, 21(1), 75-87.

[21] Kirova, V. (2021). Value co-creation and value co-destruction through interactive technology in tourism: The case of 'La Cité du Vin'wine museum, Bordeaux, France. Current Issues in Tourism, 24(5), 637-650.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 9, 2024
Accepted: Aug 17, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1474–1481, DOI 10.12694/scpe.v26i3.4274

THE INTERACTIVE SYSTEM OF MUSIC EMOTION RECOGNITION BASED ON DEEP LEARNING

JING CHEN[*]

Abstract. This paper proposes an interactive system based on spectrogram analysis, deep neural network and wavelet analysis aiming at the complexity and subjectivity of music emotion recognition. The system first uses a spectrogram to capture the time-frequency characteristics of music signals and then automatically extracts the deep emotion-related features through a convolutional neural network (CNN). This paper introduces the Mallat algorithm for wavelet decomposition to enhance the local details of audio signals to improve the accuracy of feature extraction. The experimental results show that the system performs well in recognizing music emotions, and the accuracy is significantly improved compared with the traditional method. In addition, the system supports real-time interaction, allowing users to personalize music experience by adjusting emotional labels, thus showing broad application prospects in music therapy, game entertainment and other fields. This study promotes the development of music emotion recognition technology and provides a new perspective for further exploration of deep learning in interdisciplinary applications.

Key words: Music emotion recognition; Deep learning; Spectrogram; Neural network; Wavelet analysis; Mallat algorithm

1. Introduction.

Music, as a universal and profound expression of emotions, has been closely linked to human emotions since ancient times. With the rapid development of artificial intelligence technology, especially the successful application of deep learning in image and speech recognition, music emotion recognition (MER) technology has gradually become the research focus of academia and industry. MER is designed to automatically identify and classify the emotional colors contained in musical works by analyzing musical signals, and this technological breakthrough will revolutionize music recommendation, psychotherapy, game design and many other fields.

In the music emotion recognition research process, scholars have tried many methods. Literature [1] proposes the MER method based on traditional machine learning, which solves the emotion classification problem in early music by manually extracting musical features, such as rhythm, melody, and harmony. However, this approach relies on expert knowledge and experience and is difficult to capture the nuances of musical emotion. Subsequently, literature [2] introduced deep learning technology, especially convolutional neural network (CNN). However, the traditional CNN has limitations in processing time-frequency domain information. Literature [3] uses a spectrogram as input to convert music signals into two-dimensional images so that CNN can better understand the time-frequency structure of music to overcome this problem. However, the spectrogram is insufficient to preserve the musical signal's details. Therefore, literature [4] combined wavelet analysis technology, used the Mallat algorithm to conduct multi-scale decomposition of music signals and extracted more abundant wavelet coefficient features. This feature shows superiority in describing the dynamic change of musical emotion. However, strategies to effectively combine wavelet analysis with deep learning still need further exploration.

This paper aims to study an interactive music emotion recognition system based on deep learning, which combines spectrogram, deep neural network, and wavelet analysis to realize more accurate music emotion recognition [5]. First, this paper will discuss how to use the spectrogram as the input of CNN to capture the time-frequency characteristics of music signals. Secondly, this paper will introduce the Mallat algorithm for wavelet decomposition to extract multi-scale features of music signals and combine them with the feature extraction layer of CNN to enhance the emotion recognition ability of the system. In addition, this paper

[*] Public Art Teaching Department, Zhengzhou College of Finance and Economics, Zhengzhou 450000, China (Corresponding author, cccdd1232024@163.com)

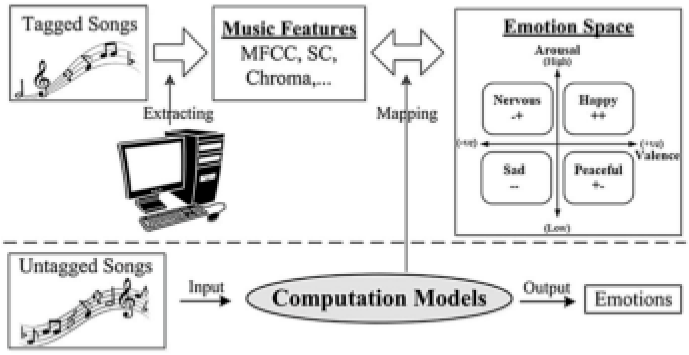


Fig. 2.1: Basic framework of music emotion recognition model.

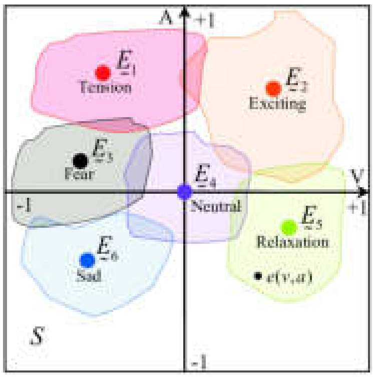


Fig. 2.2: V-A emotional space diagram.

will also study how to design a real-time interactive interface so that users can participate in the emotion recognition process and constantly optimize the system's performance through a feedback mechanism [6]. In the experimental part, this paper will collect various music data sets, including music works of different styles and emotional tendencies, to verify the generalization ability and accuracy of the system [7].

2. Deep learning neural network model.

2.1. Model Framework. This project intends to build a song emotion recognition model based on deep neural networks and machine learning technology. Figure 2.1 shows an infrastructure diagram of this pattern.

Firstly, the music library containing different emotional markers is divided into two parts: the first part is to preprocess the original score, the second part is to extract the corresponding emotional markers, and the last part is to establish the classification model with the corresponding emotional markers.

2.2. Emotional model. This paper uses Russell's Valence-Arousal model [8]. In short, effectiveness reflects two levels of emotion: positive and negative. The higher the value, the higher the positive level of emotion and the opposite negative level. The Arousal of the subject reflected the intensity of emotion. Arousal value was high, emotional intensity was high, and arousal intensity was low. The V-A emotional pattern is shown in Figure 2.2. This article will put V - A two-dimensional space-time transformation into (+ V + A), (V + A), (-v - A) (+ V - A) and so on, four different types of emotion. The corresponding results for the four types of musical emotions are given in Table 2.1.

Table 2.1: Music emotion category table.

Category	Emotion	V-A value
Emotion of the first kind	Happy	+V+A
Emotion of the second kind	Anxiety	-V+A
Emotion of the third kind	Mawkish	-V-A
Fourth emotion	Relax	+V-A

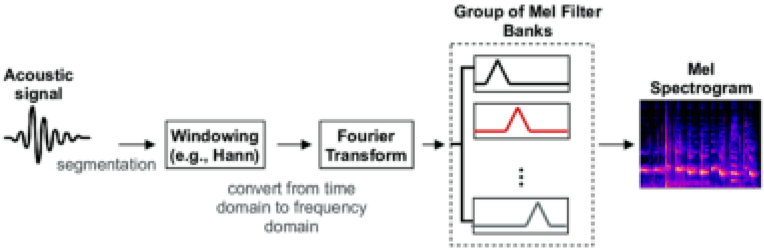


Fig. 2.3: Process of generating spectrogram.

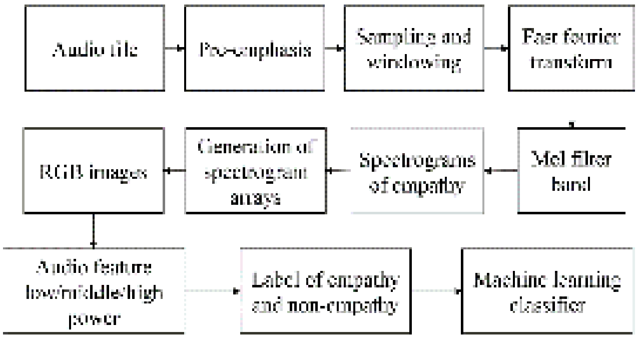

Audio file
Pre-emphasis
Sampling and windowing
Fast fourier transform
RGB images
Generation of spectrogram arrays
Spectrograms of empathy
Mel filter band
Audio feature low/middle/high power
Label of empathy and non-empathy
Machine learning classifier

Fig. 2.4: Schematic diagram of music signal generation.

2.3. Spectrogram. The spectrum is a graph obtained after Fourier analysis in the time domain. It is a two-dimensional time-frequency graph used to characterize the spectrum change horizontally and vertically. It is time horizontally and frequency vertically. The spectrum contains rich spectrum characteristics. It includes formant, energy and other frequency domain parameters and has both time and frequency domain characteristics [9]. The graph contains the entire spectrum that has not been processed, so the information about the music in the graph is not destroyed. The generation process of the spectrogram is shown in Figure 2.3.

A frame window-adding, short-time Fourier transform is performed to convert the time domain information into the frequency domain to generate the graph, and then the scale is converted into the decibel value expression of the amplitude [10]. Then, this frequency domain information is segmented and connected according to the time series to obtain the graph (Figure 2.4).

In this paper, the hearing characteristics of the human ear are taken as the primary frequency band, so the spectrum mentioned in this paper is the spectrum of the Mayer frequency band [11]. The graph takes time as the horizontal axis, Meir frequency as the vertical axis, and data energy of music signal as the coordinate. Because it is carried out in the 2D plane, its energy is represented by color, and the stronger the color, the higher the intensity of its sound.

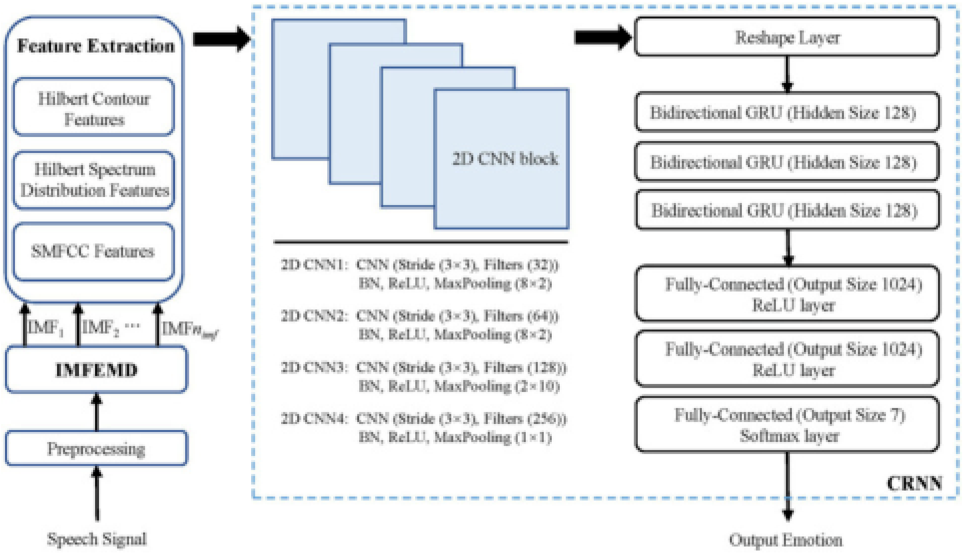

Feature Extraction
Hilbert Contour Features
Hilbert Spectrum Distribution Features
SMFCC Features
IMF_1
IMF_2 ···
$IMFn_{imf}$
IMFEMD
Preprocessing
Speech Signal
2D CNN block
2D CNN1: CNN (Stride (3×3), Filters (32))
BN, ReLU, MaxPooling (8×2)
2D CNN2: CNN (Stride (3×3), Filters (64))
BN, ReLU, MaxPooling (8×2)
2D CNN3: CNN (Stride (3×3), Filters (128))
BN, ReLU, MaxPooling (2×10)
2D CNN4: CNN (Stride (3×3), Filters (256))
BN, ReLU, MaxPooling (1×1)
Reshape Layer
Bidirectional GRU (Hidden Size 128)
Bidirectional GRU (Hidden Size 128)
Bidirectional GRU (Hidden Size 128)
Fully-Connected (Output Size 1024)
ReLU layer
Fully-Connected (Output Size 1024)
ReLU layer
Fully-Connected (Output Size 7)
Softmax layer
CRNN
Output Emotion

Fig. 2.5: Structure of CRNN music emotion recognition model.

2.4. CRNN model based on deep neural network. In this project, CNN was used to obtain the time series features of the atlas, and based on preserving the time series features of the atlas, the feature maps of the time series with complete features were obtained [12]. This gives a complete time series feature. The method takes speech as the essential information and CRNN as the learning object. It was using CRNN to learn its features, and to realize the end-to-end learning of music mood. The structure diagram of the CRNN is shown in Figure 2.5(image quoted in Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network).

The input to the network is the musical notation. In the substructure of the convolutional network, the advantages of the CNN network in 2D data are given full play, and the 1*15* N spectrum characteristic diagram is obtained by extracting spectrum information and maintaining the time series characteristics of the spectrum [13]. The core of this method is the convolutional pool processing of convolutional neural networks. By optimizing the convolution kernel, step size, layer number, etc., the frequency domain dimension of the obtained feature graph is reduced to 1. In this way, the signal's frequency domain and time characteristics are effectively fused [14]. It considers the feature extraction of spectrogram as an image Angle and the feature extraction of music signal time series Angle.

3. Feature information extraction and classification methods. This project takes the Simplified A-V emotional model as the research object and selects different types of emotions from four emotional modes (intense, happy, low, and soft). The intensity and speed of the music are very high in the intense areas. The music is more intense and faster in the happy zone [15]. The song is less intense in the soft area, and the tempo is slower. The music is less intense in the low zone, and the tempo is slower.

3.1. Feature Information.

3.1.1. Strength. The audience's grasp of the strength of the music is usually judged by the pitch and beat speed [16]. A physical quantity called the average energy is defined to quantify the intensity of music. The formula is:

$$E_t = \sum_{i=t*M}^{(t+1)*M} \frac{u_i}{M}; j, t = 0, 1, 2 \cdots \tag{3.1}$$

E_t is the short-term average energy of segment $t.u_i$ is the j pieces of music data collected, and M is the number of music data collected for each segment.

3.1.2. Rhythm. Strong and happy music usually has a faster speed, while low and soft music has a slower speed. The relative prosody method is used to replace the complicated prosody formula. 3.2 Classification Algorithm. Considering that the music emotion characteristic information is generally selected from the two aspects of high frequency and low frequency, this paper adopts the real-time method of wavelet analysis -The Mallat method

$$b_m[n] = \sum_t l[t-2n]b_{m+1}[t] \tag{3.2}$$

$$s_m[n] = \sum_t f[t-2n]b_{m+1}[t] \tag{3.3}$$

$l[t], f[t]$ is the signal string of the pulse response and the signal of the highpass signal. The wavelet analysis method transforms the signal by discrete Fourier transform, and the amplitude in the frequency domain is obtained. λ is used to represent the base frequency, and the following formula is obtained:

$$B(\lambda) = \sum_n b(n)\exp(-j\lambda n) \tag{3.4}$$

The wavelet analysis effectively identifies the music fragments with different simultaneous frequency characteristics. The identification of genetic information is combined with the identification of sound, which significantly improves detection efficiency. Wavelet transform is used to extract the feature of the spectrum table, extract the spectrum segment with the highest amplitude, and then the pronunciation time of the adjacent spectrum segments is timed. The duration of the large and small amplitude segments is found by comparing the adjacent spectrum segments to realize the rough spectrum recognition of the spectrum segments.

Figure 3.1 shows the contrast items' frequency-amplitude graph for each mixed tone. B_1 is its magnitude. Where y_2, y_3, y_4 is the triad tone contrast term with y_2, y_3, y_4 in each triad component, and the corresponding amplitude is B_2, B_3, B_4. y_5 is the contrasting item of tone, and its magnitude is B_5. The portion with a lower amplitude is not marked and can be excluded when setting the selection threshold.

$W_t = \{w_{t1}, w_{t2}, \cdots, w_{tn}\}$ is used to represent the defined sequence, where w_{ti} represents the i comments contained in the t filtered comment items. If it's a single tone, then $i = 1$. If it's n then $i = n$. The sequence $E_{W_t} = \{E_{w11}, E_{w+2}, \cdots, E_{w+n}\}$ may be qualified by addition, while E_{wit} represents the intensity of i comments included by the t comment items being filtered, $t = 1, 2, \cdots, i = 1, 2, \cdots, n$. Set the comparison coefficient to z_t and calculate it with the following equation:

$$z_t = E_{t+1}/E_t; t = 0, 1, 2, \cdots \tag{3.5}$$

E_t represents the average value of item t recorded. This comparison can be single or juxtaposed. Its formula goes like this:

$$E_t = \overline{E_{W_t}} = \sum_{i=1}^{n} E_{wti}/n; i = 1, 2, \cdots, n, t = 1, 2, \cdots \tag{3.6}$$

In tone contrast, it's A single tone $i = n = 1$ when the mean is $E_t = E_{W_t} = E_{w+1}$. At this time, the change of the adjacent sound contrast term can be determined by the value of z_t. If the value of z_t is in the closed interval $[0.6, 1.4]$, its change can be regarded as a slight change in the same roughness region [17]. When the value of z_t exceeds this interval, it can be regarded as a jump in different roughness regions. However, this contrast leads to a common phenomenon:

$$\begin{aligned} &z_1, z_2, \cdots, z_{t-1} \in [0.6, 1.4] \\ &z_t, z_{t+1}, \cdots, z_{t+n} \notin [0.6, 1.4] \\ &z_{t+n+1}, \cdots \in [0.6, 1.4] \end{aligned}$$

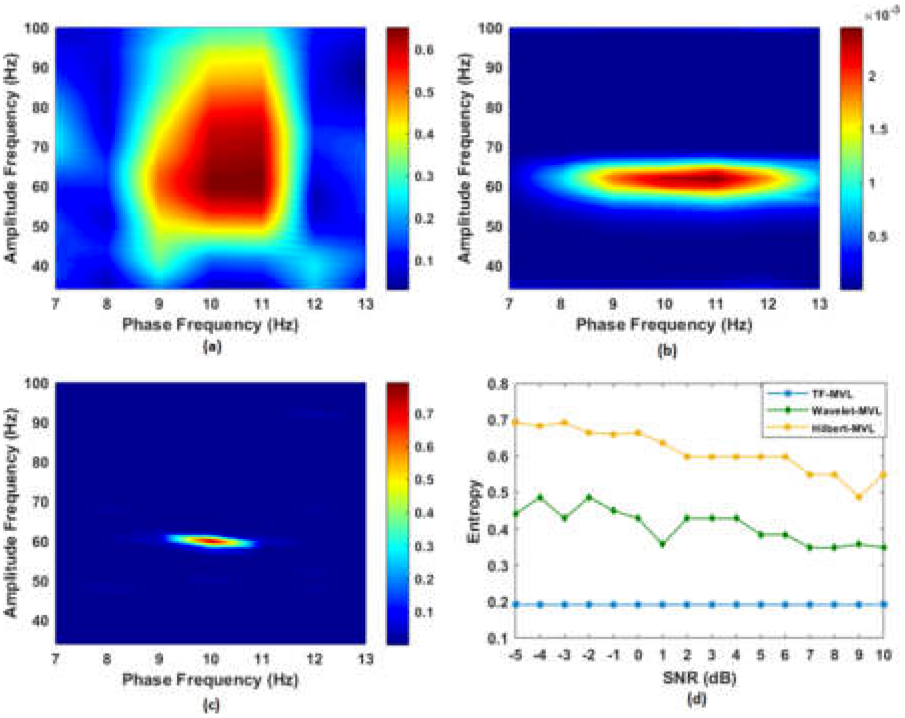


Fig. 3.1: Frequency - amplitude of the note comparison term in the mixed note bar.

The number of note comparators is M, so it only takes a simple operation to obtain a rough beat correlation value [18]. The metric-dependent value of the first paragraph is $\gamma_1 = M/t_1$. Similarly, if a song is divided into H parts, then the velocity correlation value of the segment l is

$$\gamma_l = M/t_l \tag{3.7}$$

A similar algorithm can be used for the new contrast coefficient z_t to overcome the limitation of rough partitioning based on average energy:

$$z_t = \gamma_{l+1}/\gamma_l; l = 0, 1, 2, \cdots \tag{3.8}$$

Similarly, if z_t is in a closed range [0.8,1.2], then its change can be regarded as a slight change in the same rough perception region. When z_t exceeds this interval, it can be regarded in this paper as a jump in a different roughness region.

4. Experimental results. Using the wavelet analysis software package of Matlab7.0, the rough emotional soft cutting test was carried out on the music fragments with different emotional components, which the author edited. The sampling rate is 12015 Hz. The sampling length was 50 seconds. The samples were labeled manually to determine the original emotion region [19]. In addition, a rough "soft cut" reference was made to the emotions in the test set by artificial perception in 20 researchers with good musical literacy. The results of the test are shown in Table 4.1.

Each test's maximum error time and minimum error time are 103 ms and 8 ms, respectively. The time-domain deviation of both the rough and real emotion fragments is within the acceptable range. The experiment shows that this soft-cutting technology can meet the precision requirement of the music lighting demonstration control system, and there is no apparent false connection phenomenon.

Table 4.1: Experimental results of soft cutting of music rough emotion.

Coarse affective domain	Quantity	Fall into this category	Correct number	Precision/%	Recall/%
fierce	21	23	19	85.21	94
Cheerful and cheerful	21	25	18	73.75	89
gentle	21	19	15	81.04	73
low	21	17	16	97.71	78

5. Conclusion. This paper presents an interactive system to address the challenge of music emotion recognition, which cleverly combines deep learning techniques with music signal processing. By using a spectrogram as the input of a deep neural network, the system can effectively capture the time-frequency characteristics of music, and the introduction of wavelet analysis and the Mallat algorithm further enhances the precision of feature extraction, especially in the processing of subtle changes in music emotion. The experimental results show that the designed system achieves high accuracy in the music emotion recognition task, proving the strong potential of deep learning in music emotion analysis. In addition, the interactive design of the system allows users to participate in the emotion recognition process, and the real-time feedback mechanism not only improves the user experience but also provides the possibility for continuous learning and optimization of the system.

REFERENCES

[1] Jingjing, W. A. N. G., & Ru, H. U. A. N. G. (2022). Music emotion recognition based on the broad and deep learning network. Journal of East China University of Science and Technology, 48(3), 373-380.

[2] Gómez-Cañón, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y. H., & Gómez, E. (2021). Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine, 38(6), 106-114.

[3] Xu, L., Wen, X., Shi, J., Li, S., Xiao, Y., Wan, Q., & Qian, X. (2021). Effects of individual factors on perceived emotion and felt emotion of music: based on machine learning methods. Psychology of Music, 49(5), 1069-1087.

[4] Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. Multimedia Tools and Applications, 80(2), 2887-2905.

[5] Sarkar, R., Choudhury, S., Dutta, S., Roy, A., & Saha, S. K. (2020). Recognition of emotion in music based on deep convolutional neural network. Multimedia Tools and Applications, 79(1), 765-783.

[6] Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S. (2021). Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends, 2(01), 73-79.

[7] Nawaz, R., Cheah, K. H., Nisar, H., & Yap, V. V. (2020). Comparison of different feature extraction methods for EEG-based emotion recognition. Biocybernetics and Biomedical Engineering, 40(3), 910-926.

[8] Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1), 45-55.

[9] Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. Journal of Artificial Intelligence and Systems, 2(1), 53-79.

[10] Veltmeijer, E. A., Gerritsen, C., & Hindriks, K. V. (2021). Automatic emotion recognition for groups: a review. IEEE Transactions on Affective Computing, 14(1), 89-107.

[11] Medina, Y. O., Beltrán, J. R., & Baldassarri, S. (2022). Emotional classification of music using neural networks with the MediaEval dataset. Personal and Ubiquitous Computing, 26(4), 1237-1249.

[12] Zepf, S., Hernandez, J., Schmitt, A., Minker, W., & Picard, R. W. (2020). Driver emotion recognition for intelligent vehicles: A survey. ACM Computing Surveys (CSUR), 53(3), 1-30.

[13] Schlegel, K., Palese, T., Mast, M. S., Rammsayer, T. H., Hall, J. A., & Murphy, N. A. (2020). A meta-analysis of the relationship between emotion recognition ability and intelligence. Cognition and emotion, 34(2), 329-351.

[14] Xu, G., Guo, W., & Wang, Y. (2023). Subject-independent EEG emotion recognition with hybrid spatio-temporal GRU-Conv architecture. Medical & Biological Engineering & Computing, 61(1), 61-73.

[15] Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 14(2), 715-729.

[16] Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. IEEE Signal Processing Magazine, 38(6), 59-73.

[17] Ding, Y., Robinson, N., Zhang, S., Zeng, Q., & Guan, C. (2022). Tsception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. IEEE Transactions on Affective Computing, 14(3), 2238-2250.

[18] Kamble, K. S., & Sengupta, J. (2021). Ensemble machine learning-based affective computing for emotion recognition using dual-decomposed EEG signals. IEEE Sensors Journal, 22(3), 2496-2507.

[19] Panda, R., Malheiro, R., & Paiva, R. P. (2020). Audio features for music emotion recognition: a survey. IEEE Transactions on Affective Computing, 14(1), 68-88.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 10, 2024
Accepted: Aug 17, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1482–1488, DOI 10.12694/scpe.v26i3.4323

APPLICATION OF DATA VISUALIZATION INTERACTION TECHNOLOGY IN AEROSPACE DATA PROCESSING

TIANFENG LI *

Abstract. In the aerospace sector, efficient data processing is critical to ensuring flight safety and improving operational efficiency. Firstly, an aircraft 3D modeling method based on OpenGL technology is introduced. This method realizes highly realistic aircraft models through accurate geometric rendering and material mapping. At the same time, the Bursa-Wolf method is used for coordinate transformation to ensure the accuracy and consistency of the model from different perspectives. Then, this paper discusses the application of visual interaction technology in aerospace data processing, especially in-flight data visualization systems. The simulation results show that the system can receive and process a lot of flight data in real-time and display the aircraft's attitude, trajectory, and critical parameters through an intuitive graphical interface so that pilots and ground controllers can make decisions quickly. This technology improves the efficiency of data processing and enhances the comprehensibility and usability of data. The stability and reliability of the technology in complex environments are verified by simulating actual flight scenarios.

Key words: Aircraft 3D model modeling method; Visual interaction technology; Space data; Flight data visualization system; OpenGL technology; Bursa-Wolf method

1. Introduction. Data processing is increasingly required in the aerospace sector to ensure flight safety, improve operational efficiency, and advance scientific innovation. Data visualization interaction technology is crucial as a bridge between raw data and human cognition. Regarding aircraft 3D model modeling methods, early studies mainly rely on manual modeling and simple geometric transformation. Although this method can meet the needs of static display to a certain extent, it is challenging to deal with complex dynamic flight data. With the development of computer graphics, OpenGL technology has gradually become the mainstream tool for aircraft 3D modeling. Literature [1] proposes an aircraft 3D modeling method based on OpenGL, which realizes a highly realistic aircraft model through fine geometric rendering and material mapping and solves the shortcomings of traditional modeling methods in detail presentation and dynamic display. Regarding visual interaction technology, traditional data processing methods are often limited to two-dimensional planes, and it is difficult to show complex spatial relationships and dynamic changes intuitively. Literature [2] applies visual interaction technology to space data processing, realizing real-time tracking and analysis of complex spacecraft motion trajectories and providing powerful decision support for ground controllers.

Regarding flight data visualization systems, a sound system needs to process large amounts of real-time data and be able to present this data intuitively and understandably. The researchers developed flight data visualization systems. For example, the system designed in the literature [3] can receive and process data from aircraft sensors in real-time and display the aircraft's attitude, trajectory, and critical parameters through an intuitive graphical interface. These systems increase data processing automation and reduce the risk of human error. Reference [4] describes in detail the application of the Bursa-Wolf method in aircraft 3D model modeling, which ensures the accuracy and consistency of the model under different perspectives through accurate coordinate conversion. The application of this method makes the establishment of the aircraft model more accurate and provides a solid foundation for the subsequent visual interaction.

Data visualization and interaction technology have broad application prospects and essential practical value in aerospace data processing. This paper will start with the three-dimensional modeling method of aircraft, introduce the application of visual interaction technology in-flight data visualization systems in detail, and discuss its potential to improve data processing efficiency, ensure flight safety and promote scientific research

[*] NanYang Institute of Technology, Nanyang, Henan, 473004, China (Corresponding author, longyust_001@163.com)

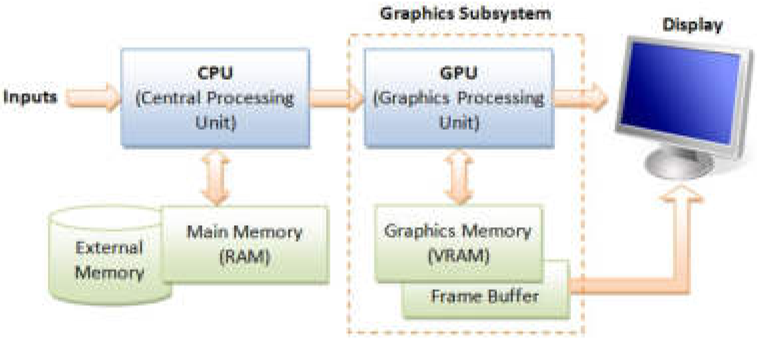


Fig. 2.1: How OpenGL works in a Windows environment.

and innovation [5]. Through system simulation and actual case analysis, this paper will verify the feasibility and superiority of the proposed technology and provide helpful references for future research and practice.

2. Open graphics library architecture for Windows. OpenGL, an open-source 3D drawing tool, can work closely with Visual C++. In this way, the related and drawing operations are completed, ensuring the method's effectiveness and reliability [6]. OpenGL adopts a client /Server approach to implement, which provides a good solution for OpenGL. OpenGL's graphics library is wrapped in OpenGL 32.DLL. When used by the client, all OpenGL functions are processed by OPENGL32.DLL and then sent to the server. The OpenGL instructions are reprocessed and sent directly to the Win32 device driver interface to send the finished image instructions to the image display driver. This process is shown in Figure 2.1.

3. The architecture development of OpenGL based on MFC.

3.1. OpenGL implementation method on MFC. OpenGL does not support Windows management, performed on a platform-specific system. It is necessary first to connect Windows Visual with OpenGL to realize 3D visual design based on OpenGL language. The 3D image is rendered and processed by OpenGL language [7]. It is necessary to build a simulated aircraft model and then bundle it with a modeled dialog box to realize OpenGL graphics in the MFC environment. OpenGL allows 3D images to be drawn by "rendering context."

3.2. Dual Buffer technology. The simulation window is constantly updated and rendered to generate an animation. By default, the image will flash. There are two reasons for the flash: one is the background elimination, and the other is the rendering time is too long. The dual cache mechanism is used in OpenGL. Partition the two frame caches. When a 3D surface appears consecutively, the data in one of the frame caches is surface-rendered. In the other frame, the cache is used for image processing [8]. When a video in the background cache needs to be displayed, OpenGL copies it to the front-end cache. The cache continuously reads the information in the cache and outputs it to the screen. Due to dual caching technology, each 3D surface does not appear after the end of rendering, so the viewer can directly view each 3D image.

4. Three-dimensional modeling of aircraft model.. The 3D modeling of aircraft is an integral part of its visualization process. The structure, materials, positioning and other information of the vehicle are contained in it. Due to OpenGL's lack of advanced 3D modeling instructions, it is challenging to generate a complex aircraft model [9] programmatically. The aircraft was modeled with high precision by 3D modeling software and read by OpenGL. Because of the different 3D modeling documents, their data can be read differently. This makes it more difficult for developers. Therefore, this paper proposes an easy-to-read intermediate format document to realize the conversion between various 3D modeling documents. When modeling OpenGL, the MilkShape3D format was selected. The complexity of modeling and the need for computer hardware should be reduced as much as possible to improve the fluency of rendering. The workflow of actual 3D model construction is shown in Figure 4.1.

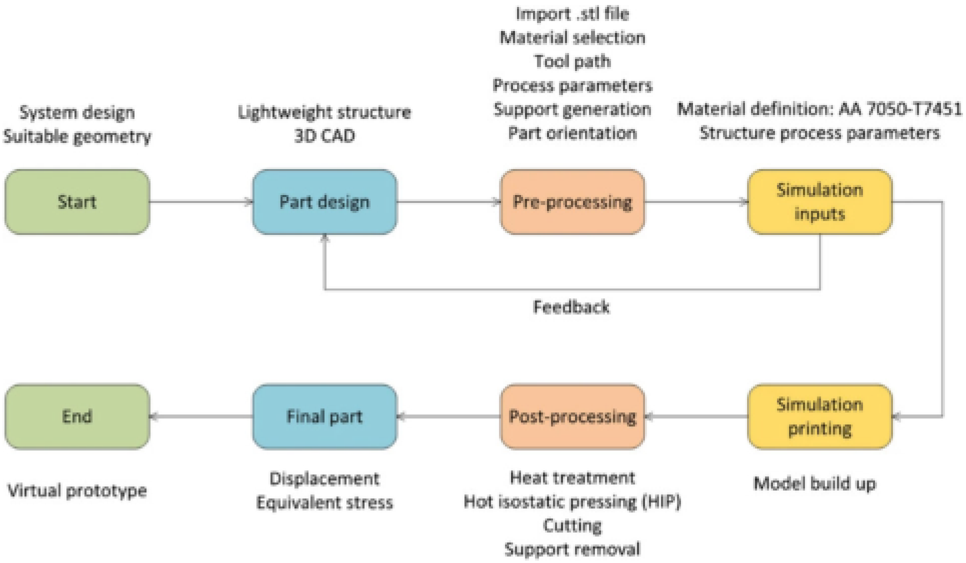


Fig. 4.1: Aircraft 3D model making process.

4.1. Virtual environment rendering. The visualization of "flight history" mainly consists of the following aspects: the appearance and texture of the aircraft as an external document. The aircraft model file is automatically matched according to the aircraft type information in the flight measurement data attribute. For example, if there is A "Test Object Name= A" in the aircraft measurement data description file, the AMS3D mode will be automatically found and loaded when the data is replayed. The milk shape Model was used to encapsulate the 3D modeling of the aircraft. Milk shape Model: Model Data () can load the aircraft model from the MS3D file and then convert it into a point, polygon, and material list. Draw () renders the aircraft. Since flight trajectory data display technology focuses on the realistic restoration of aircraft attitude, maneuvering and other motions, the simulation degree of the ground environment is low, so it is transformed into ground texture mapping [10]. The results were excellent based on the spherical dome's diameter, textured with a perfect picture of the clouds. The new wing target model generates a column of smoke of arbitrary length, which is used to identify the target's trajectory.

4.2. Visualization of flight data based on measured data. This project uses aircraft attitude simulation as the primary research method to study the relationship between aircraft attitudes. Because some models do not have autonomous navigation and positioning equipment, they must be determined by comprehensive calculation of various parameters in the movement process [11]. The trajectory is solved by integrating the measured aircraft altitude, velocity, Angle of attack, pitch Angle, pitch Angle and yaw Angle. The visualization software of the flight process is designed, and the real-time playback of aircraft measurement parameters is realized. The Flight simulation view starts timing, generating a user zone failure message at a rate of 30 times/second to trigger an upgrade simulation view. The simulation window and the time domain curve frame are loosely coupled by an information-driven method to realize simple and independent programming logic. This project proposes a method based on Windows message queue accumulation information to realize non-synchronous flight data updates. To solve the problem of system response stagnation caused by redrawing due to carrier attitude change [12]. The aircraft trajectory information display system restores the aircraft attitude and the display screen of the aircraft instrument. The tracking view is also known as the wingman view, that is, the tracking view of the posture of the aircraft from the outside of the aircraft. In the process of observation, the moving center of gravity of the aircraft is taken as the starting point of its relative coordinate system, and it is fixed with the relative position of the aircraft [13]. The viewing Angle is always consistent with the vehicle's center of gravity. Its Angle to the X and Z axes and distance to the starting point. Using ordinary input and output devices, the user can adjust the position of the observation point at any Angle to realize the attitude observation of the aircraft.

5. Six degrees of freedom simulation of aircraft.. The aircraft's motion in the three-dimensional space is a complex action with many degrees of freedom, and its operating state can be obtained through the actual measurement of the aircraft [14]. Data include GPS longitude, latitude, altitude, flight speed, pitch Angle, yaw Angle, etc. In the visual display of the aircraft, it is necessary to consider the change of six parameters, such as the orientation of the center of gravity of the aircraft (x, y, z) and, the pitch angle of the aircraft and the yaw Angle. Here, the yaw Angle ψ is the tilt Angle of the vertical axis of the aircraft in the horizontal plane and the predetermined course, the pitch Angle γ is the Angle from the fuselage coordinate system x to the horizontal plane, and the yaw Angle θ is the Angle from the fuselage coordinate system y to the vertical surface passing through the z axis.

The 6-dimensional pose change, yaw Angle ψ, pitch Angle γ and roll Angle θ of the aircraft can be obtained using the rotation and shift transformation of the solid geometric coordinate system [15]. The goal of the shift transformation in this process is to convert the starting point of the local coordinate system to (x, y, z). That is, without changing the orientation and dimensions of the defined object, the components of each vertex coordinate system in the scene are represented as x, y and $z.\lambda(x, y, z, 1)$ is the uniform coordinate of any point in space, and S_x, S_y, S_z is the amount that moves along the x, y, and z axes. The change matrix for this operation is as follows:

$$\lambda'\left(x', y', z', 1\right) = \lambda(x, y, z, 1) \cdot \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & y \\ 0 & 0 & 1 & z \\ S_x & S_y & S_z & 1 \end{bmatrix} = R\left(S_x, S_y, S_z\right) \tag{5.1}$$

The three-dimensional rotation transformation is the object's rotation around its axis, and the right-hand rule determines its rotation direction. After rotation, the size and shape of the object itself does not change, but changes its position [16]. If $\lambda(x, y, z, 1)$ is the uniform coordinate of any point in space and β is the Angle of rotation around the x, y, and z axes, then the equation for rotation around the x, y, and z axes is:

$$\lambda'\left(x', y', z', 1\right) = \lambda(x, y, z, 1) \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\beta & \sin\beta & 0 \\ 0 & -\sin\beta & -\cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = Q_{x(A)} \tag{5.2}$$

$$\lambda'\left(x', y', z', 1\right) = \lambda(x, y, z, 1) \cdot \begin{bmatrix} \cos\beta & \sin\beta & 0 & 0 \\ -\sin\beta & \cos\beta & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = Q_{z(A)} \tag{5.3}$$

$$\lambda'\left(x', y', z', 1\right) = \lambda(x, y, z, 1) \cdot \begin{bmatrix} -\cos\beta & 0 & -\sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ \sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = Q_{r(F)} \tag{5.4}$$

KML uses WGS-84 as the benchmark. The Z-axis is the conventional polar CTP direction specified by BIH1984.o. The X -axis is the zero-meridian plane of BIH1984.o intersecting the CTP equator. The Y, Z and X axes form a right-handed coordinate system [17]. The actual flight track data adopts the national coordinate system of the NMEA-o183 standard. This article will use the Bursha-Wolf transformation method, which includes 7 parameters:

$$\begin{bmatrix} X_\mu \\ Y_\mu \\ Z_\mu \end{bmatrix} = \begin{bmatrix} 1 & \zeta_Z & -\zeta_Y \\ -\zeta_Z & 1 & \zeta_X \\ -\zeta_Y & -\zeta_X & 1 \end{bmatrix} \begin{bmatrix} X_L \\ Y_L \\ Z_L \end{bmatrix} + (1+\varphi) \times \begin{bmatrix} X_L \\ Y_L \\ Z_L \end{bmatrix} + \begin{bmatrix} \Delta X_0 \\ \Delta Y_0 \\ \Delta Z_0 \end{bmatrix} \tag{5.5}$$

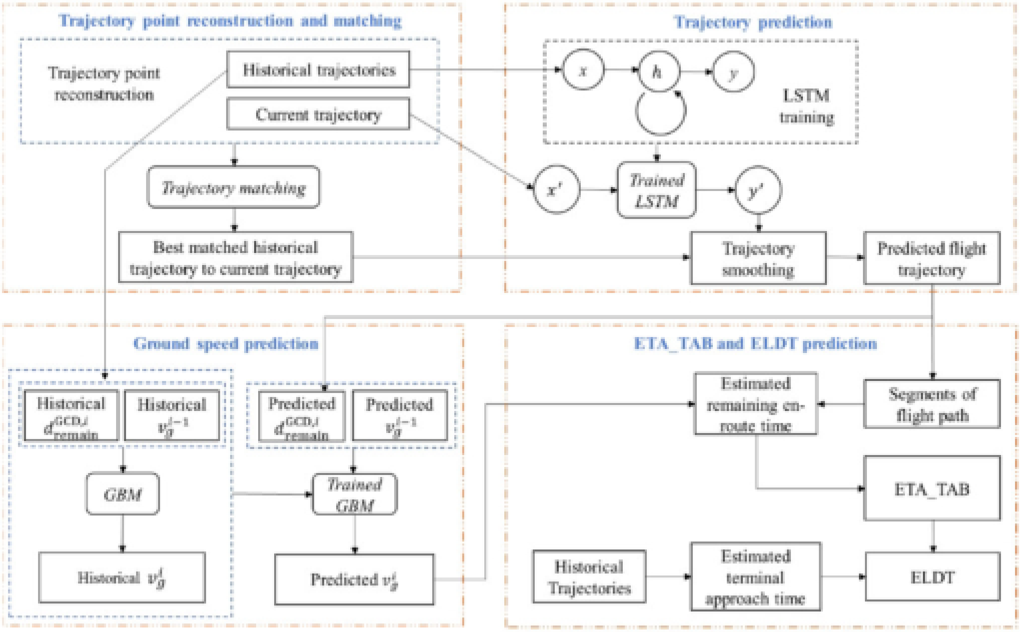

Trajectory point reconstruction and matching
Trajectory point reconstruction
Historical trajectories
Current trajectory
Trajectory matching
Best matched historical trajectory to current trajectory
Trajectory prediction
x
h
y
LSTM training
x'
Trained LSTM
y'
Trajectory smoothing
Predicted flight trajectory
Ground speed prediction
Historical $d^{GCD,i}_{remain}$
Historical v^{i-1}_g
GBM
Historical v^i_g
Predicted $d^{GCD,i}_{remain}$
Predicted v^{i-1}_g
Trained GBM
Predicted v^i_g
ETA_TAB and ELDT prediction
Estimated remaining en-route time
Segments of flight path
ETA_TAB
Historical Trajectories
Estimated terminal approach time
ELDT

Fig. 5.1: Flight trajectory data preprocessing process.

$\Delta X_0, \Delta Y_0, \Delta Z_0$ represents the translational coefficient, $\zeta_X, \zeta_Y, \zeta_Z$ represents the rotational coefficient, and φ represents the proportional coefficient. The formula (5.5) can be further transformed into

$$\begin{bmatrix} X_\mu \\ Y_\mu \\ Z_\mu \end{bmatrix} = \begin{bmatrix} X_L \\ Y_L \\ Z_L \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & -Z_L & Y_L & X_L \\ 0 & 1 & 0 & Z_L & 0 & -X_L & Y_L \\ 0 & 0 & 1 & -Y_L & X_L & 0 & Z_L \end{bmatrix} \begin{bmatrix} \Delta X_0 \\ \Delta Y_0 \\ \Delta Z_0 \\ \zeta_X \\ \zeta_Y \\ \zeta_Z \\ \varphi \end{bmatrix} \tag{5.6}$$

Seven parameter values are obtained by using more than 3 known points. Bursa-Wolf equation is used to calculate the position of the point to be measured, and then the coordinates meeting the requirements of KML specification are calculated according to the coordinates of each point to be measured. A trajectory data processing component based on KML is proposed [18]. This project intends to take the observation data of the NMEA-0183 satellite as the research object, use the Bursa-Wolf algorithm to carry out spatial transformation and achieve high-precision transformation from space to KML. This project intends to analyze the space orbit data obtained from the aerial survey in depth. With the expression of high-definition images, such as Google Earth, it is applied to the aerospace field [19]. Visualizing massive abstract data in aerospace lays a solid theoretical and practical foundation for research and application in related fields. The overall preprocessing process of flight trajectory data is shown in Figure 5.1 (the picture is quoted in Aerospace 2023, 10(8), 675).

6. Implementation and verification of the system. This paper presents a visualization method of aircraft attitude simulation based on OpenGL and transforms it into KML format. It takes full advantage of Google Earth's high-resolution images to visualize them in 3D. Finally, each function module is embedded in the measurement data management platform system. They use a unified standard system, database structure, data storage and transmission structure to build the measurement data storage, management, query, and later data processing. The software uses B/S and C/S combined architecture to establish an efficient data storage

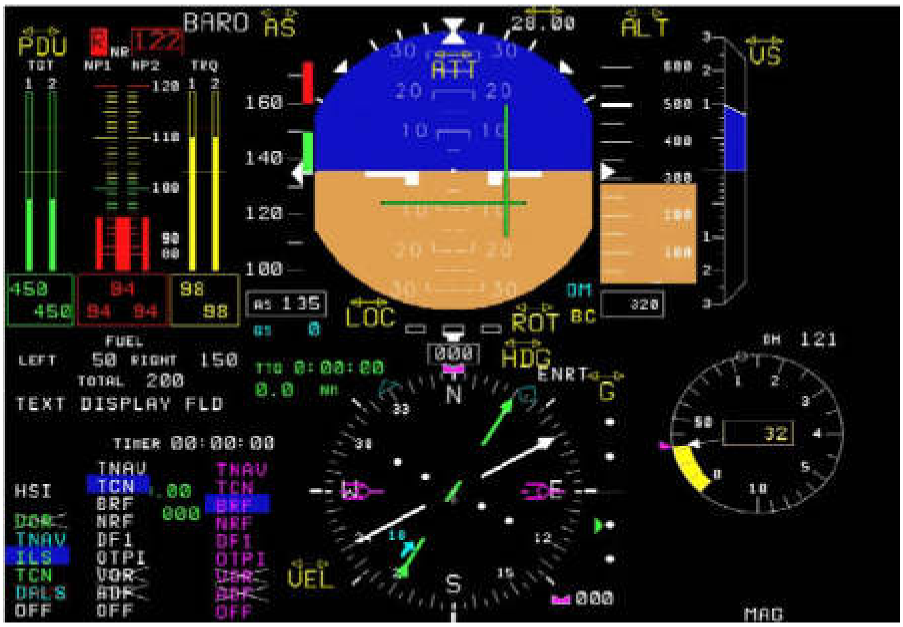


Fig. 6.1: *Visual display of single take-off and landing flight.*

system. At the same time, the 3D visualization and interactive processing of measured data are realized by using the display based on C/S structure. The results are shown in Figure 6.1.

7. Conclusion. This paper profoundly studies the application of data visualization and interaction technology in aerospace data processing, focusing on aircraft 3D model modeling methods, visual interaction technology, flight data visualization systems, and related OpenGL technology and the Bursa-Wolf method. The highly realistic aircraft model display is realized through the fine modeling of the aircraft 3D model, combined with OpenGL technology. At the same time, the Bursa-Wolf method is used for accurate coordinate transformation to ensure the consistency and accuracy of the model from different perspectives. The flight data visualization system designed in this paper can receive and process a large amount of flight data in real time and display aircraft attitude, trajectory, and critical parameters through an intuitive graphical interface, significantly improving the efficiency and accuracy of data processing. The simulation results show that the technology can quickly respond to the changes in flight data and effectively assist pilots and ground controllers in making decisions. However, although the existing technology has made significant progress, there are still some challenges and limitations. For example, how to further improve the real-time and interactivity of data visualization to support the dynamic decision-making process better is still a problem worthy of in-depth study. In addition, with the advent of the significant data era, how to effectively process and analyze larger data sets is also an important direction for future research. In future work, this paper looks forward to further optimizing the performance and functionality of data visualization interaction technology by introducing more advanced algorithms and technologies. At the same time, this paper also hopes to better integrate data visualization interaction technology into all aspects of aerospace data processing through interdisciplinary cooperation and research, thereby promoting the development and progress of the entire field. Although the current research has achieved specific results, the application of data visualization interaction technology in aerospace data processing will have more excellent development space and broader application prospects in the future.

REFERENCES

[1] Tadeja, S. K., Seshadri, P., & Kristensson, P. O. (2020). AeroVR: An immersive visualisation system for aerospace design and digital twinning in virtual reality. The Aeronautical Journal, 124(1280), 1615-1635.

[2] Dimara, E., Zhang, H., Tory, M., & Franconeri, S. (2021). The unmet data visualization needs of decision makers within organizations. IEEE Transactions on Visualization and Computer Graphics, 28(12), 4101-4112.

[3] Su, S., Perry, V., Bravo, L., Kase, S., Roy, H., Cox, K., & Dasari, V. R. (2020). Virtual and augmented reality applications to support data analysis and assessment of science and engineering. Computing in Science & Engineering, 22(3), 27-39.

[4] Chinchanikar, S., & Shaikh, A. A. (2022). A review on machine learning, big data analytics, and design for additive manufacturing for aerospace applications. Journal of Materials Engineering and Performance, 31(8), 6112-6130.

[5] Korkut, E. H., & Surer, E. (2023). Visualization in virtual reality: a systematic review. Virtual Reality, 27(2), 1447-1480.

[6] Ning, S., Sun, J., Liu, C., & Yi, Y. (2021). Applications of deep learning in big data analytics for aircraft complex system anomaly detection. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 235(5), 923-940.

[7] Borgen, K. B., Ropp, T. D., & Weldon, W. T. (2021). Assessment of augmented reality technology's impact on speed of learning and task performance in aeronautical engineering technology education. The International Journal of Aerospace Psychology, 31(3), 219-229.

[8] Xiong, M., & Wang, H. (2022). Digital twin applications in aviation industry: A review. The International Journal of Advanced Manufacturing Technology, 121(9), 5677-5692.

[9] Fang, X., Wang, H., Liu, G., Tian, X., Ding, G., & Zhang, H. (2022). Industry application of digital twin: From concept to implementation. The International Journal of Advanced Manufacturing Technology, 121(7), 4289-4312.

[10] Munir, A., Blasch, E., Kwon, J., Kong, J., & Aved, A. (2021). Artificial intelligence and data fusion at the edge. IEEE Aerospace and Electronic Systems Magazine, 36(7), 62-78.

[11] Fan, W., Fu, Q., Cao, Y., Zheng, L., Zhang, X., & Zhang, J. (2024). Binocular vision and priori data based intelligent pose measurement method of large aerospace cylindrical components. Journal of Intelligent Manufacturing, 35(5), 2137-2159.

[12] Guo, H., Liang, D., Chen, F., & Shirazi, Z. (2021). Innovative approaches to the sustainable development goals using Big Earth Data. Big Earth Data, 5(3), 263-276.

[13] Solmaz, S., & Van Gerven, T. (2022). Automated integration of extract-based CFD results with AR/VR in engineering education for practitioners. Multimedia Tools and Applications, 81(11), 14869-14891.

[14] Hu, W., Zhang, T., Deng, X., Liu, Z., & Tan, J. (2021). Digital twin: A state-of-the-art review of its enabling technologies, applications and challenges. Journal of Intelligent Manufacturing and Special Equipment, 2(1), 1-34.

[15] Lv, Z., Qiao, L., Cai, K., & Wang, Q. (2020). Big data analysis technology for electric vehicle networks in smart cities. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1807-1816.

[16] Rawat, D. B., & El Alami, H. (2023). Metaverse: Requirements, architecture, standards, status, challenges, and perspectives. IEEE Internet of Things Magazine, 6(1), 14-18.

[17] Malekloo, A., Ozer, E., AlHamaydeh, M., & Girolami, M. (2022). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. Structural Health Monitoring, 21(4), 1906-1955.

[18] McStraw, T. C., Pulla, S. T., Jones, N. L., Williams, G. P., David, C. H., Nelson, J. E., & Ames, D. P. (2022). An Open-Source Web Application for Regional Analysis of GRACE Groundwater Data and Engaging Stakeholders in Groundwater Management. JAWRA Journal of the American Water Resources Association, 58(6), 1002-1016.

[19] Hasan, S. M., Lee, K., Moon, D., Kwon, S., Jinwoo, S., & Lee, S. (2022). Augmented reality and digital twin system for interaction with construction machinery. Journal of Asian Architecture and Building Engineering, 21(2), 564-574.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 22, 2024
Accepted: Aug 19, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1489–1495, DOI 10.12694/scpe.v26i3.4324

THE DESIGN AND TESTING OF INTELLIGENT ORCHARD PICKING SYSTEM FOR AGRICULTURAL MACHINERY BASED ON IMAGE PROCESSING TECHNOLOGY

GUIMING QIAN*

Abstract. This paper proposes an intelligent orchard-picking system for agricultural machinery based on image processing technology. The system uses binocular vision technology to obtain high-precision 3D point cloud data of fruit trees and uses the Mask RCNN algorithm to detect and segment fruit. The system design includes two parts: hardware selection and software algorithm implementation. The hardware part mainly includes a binocular camera, robot arm and end actuator, while the software part integrates image preprocessing, target recognition and positioning, path planning, and grasp control. In the system simulation stage, the whole process is optimized several times to ensure its stability and reliability in practical application. Finally, the effectiveness and practicability of the system are verified by testing in a natural orchard environment. The experimental results show that the system can accurately identify and locate fruits under complex backgrounds, realize efficient and automatic picking operations, and significantly improve orchards' production efficiency and economic benefits. The research results of this paper are of great significance for promoting the intelligent development of agricultural machinery.

Key words: Image processing technology; Intelligent orchard picking system; Binocular vision; Mask RCNN algorithm; System simulation

1. Introduction. Image processing technology is vital in intelligent agricultural machinery as an essential branch of computer vision. The fruit can be recognized and positioned efficiently and accurately through image acquisition, preprocessing, feature extraction and classification recognition.

In the design of intelligent orchard-picking systems, binocular vision technology has been widely concerned because of its ability to provide three-dimensional spatial information. Through binocular vision technology, the system can obtain the depth information of the fruit, to achieve an accurate grasp of the fruit. At the same time, combined with advanced deep learning algorithms such as Mask RCNN, the system can further improve the fruit's recognition rate and positioning accuracy. Literature [1] proposes a fruit-picking robot system based on image processing and machine vision. The system realizes the recognition and positioning of fruits through image processing technology and uses machine vision technology to guide robots to carry out picking operations. In reference [2], a fruit recognition method based on deep learning is proposed to solve the problem of fruit recognition under a complex background. This method uses a convolutional neural network to extract and classify fruit images, effectively improving fruit recognition rate. In addition, literature [3] also studied the 3D localization technology of fruit based on binocular vision. The three-dimensional coordinate information of fruit was obtained by binocular vision technology, which provided accurate location information for subsequent picking operations.

However, although predecessors have made substantial progress in this field, some problems still need to be solved. For example, in the complex and changing orchard environment, achieving fast and accurate identification and positioning of fruits remains a challenge. At the same time, how to improve the robustness and adaptability of the system so that it can adapt to different varieties, sizes and colors of fruit is also the hot and challenging point of current research.

An intelligent orchard-picking system for agricultural machinery based on image processing technology is proposed in this paper. The system integrates binocular vision technology and Mask RCNN algorithm to realize fast and accurate recognition and localization of fruit [4]. First, the fruit's depth information and three-dimensional coordinate information were obtained by binocular vision technology. Then, the Mask RCNN algorithm segmented and identified the fruit finely. Finally, according to the recognition results, the robot

*Sichuan Vocational and Technical College, Suining Sichuan 629000, China (Corresponding author, 18190203339@163.com)

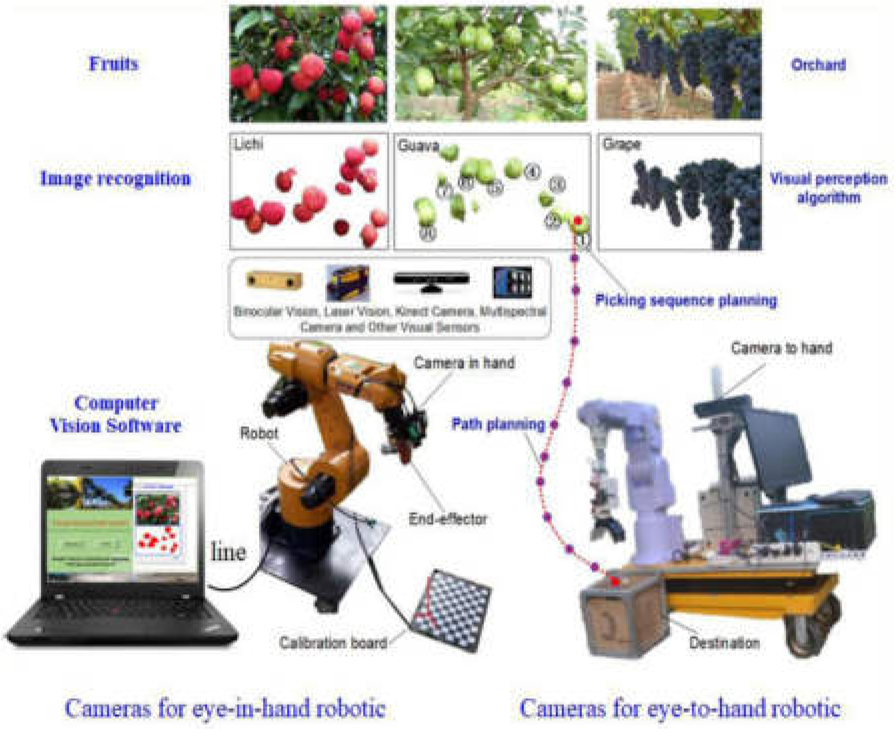

Fruits
Orchard
Image recognition
Lichi
Guava
Grape
Visual perception algorithm
Binocular Vision, Laser Vision, Kinect Camera, Multispectral Camera and Other Visual Sensors
Picking sequence planning
Camera to hand
Camera in hand
Computer Vision Software
Robot
Path planning
End-effector
line
Calibration board
Destination
Cameras for eye-in-hand robotic
Cameras for eye-to-hand robotic

Fig. 2.1: Architecture of agricultural intelligent orchard picking system.

arm is controlled to carry out the automatic picking operation. The research in this paper not only helps to promote the development of intelligent agricultural machinery and improve the efficiency and quality of orchard production but also provides a valuable reference for researchers and engineers in related fields.

2. Design of agricultural intelligent orchard picking system based on image processing.

2.1. System Architecture.. The intelligent orchard-picking system designed in this study adopts the modular design concept, and the overall architecture is divided into four main parts: image acquisition, data processing, decision control and executive mechanism. The image acquisition module is responsible for the real-time capture of orchard scene images, using high-resolution cameras and binocular vision technology to obtain two-dimensional images and three-dimensional spatial information of fruits [5]. After receiving the image information, the data processing module uses image processing technology and deep learning algorithms, such as Mask RCNN, for image preprocessing, feature extraction, fruit recognition and location. The decision control module generates corresponding control instructions according to the processing results, coordinates the actions of the robotic arm and the end effector, and realizes the precise picking of the fruit. The actuator module comprises a robotic arm and an end actuator responsible for specific picking operations [6]. The whole system realizes the data transmission and cooperative work between modules through wireless communication technology to ensure the system's efficient operation. Figure 2.1 shows the architecture of the agricultural intelligent orchard Picking system (the picture is quoted in Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review).

2.2. Hardware Design. The system mainly includes image acquisition equipment, a robot arm system, an end effector, and a central control unit. The image acquisition device consists of a high-resolution RGB camera and an infrared camera to obtain the color and depth information of the fruit [7]. The RGB camera captures the appearance characteristics of the fruit, and the infrared camera provides the depth information of the fruit by measuring the difference in thermal radiation on the surface of the fruit. Combining the two can

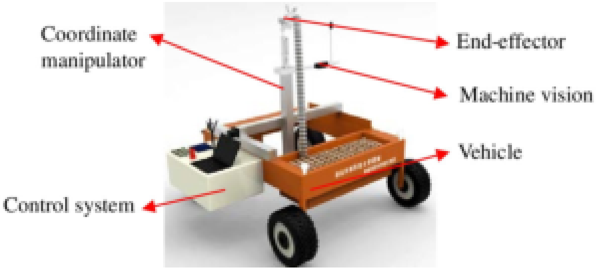


Fig. 2.2: Hardware structure diagram of fruit picking robot.

achieve an accurate three-dimensional reconstruction of the fruit [8]. The industrial robot arm with multiple degrees of freedom is selected for the robot arm system, which has flexible movement ability and high positioning accuracy and can adapt to the complex and changeable orchard environment. The end-effector is designed with an adjustable claw structure that ADAPTS the shape and size of the fruit to ensure stability and reliability during the picking process [9]. The central control unit adopts a high-performance embedded processor, which is responsible for the overall control and data processing of the system and ensures the coordinated operation of each module. Figure 2.2 shows the hardware structure diagram of the fruit-picking robot.

2.3. Control system and software design. The design of the control system and software is the key to realizing the function of an intelligent orchard-picking system. The control system adopts a hierarchical design, including the bottom controller and the upper computer control software [10]. The bottom controller receives instructions from the central control unit and drives the robot arm and the end actuator to perform the corresponding actions. The upper computer control software runs on the high-performance computer, displays the system's working state and parameter Settings through the graphical user interface (GUI), and provides the user interaction interface. Regarding software design, the system adopts the modular programming idea. The image processing, decision control and motion planning function modules are developed independently, which makes it easy to maintain and upgrade the system [11]. The image processing module uses the deep learning framework TensorFlow and PyTorch to realize the training and deployment of the Mask RCNN algorithm and improve fruit recognition rate and positioning accuracy. According to the image processing results and path planning algorithm, the decision control module generates the optimal picking path and action sequence to ensure that the robot arm can complete the picking task efficiently and safely [12]. The structure diagram of the Robot control System is shown in Figure 2.3 (the picture is quoted in the Four-wheeled Mobile Robot with Autonomous Navigation System in ROS). In software implementation, special attention is paid to the real-time and robustness of the system. Through multi-thread technology and a real-time operating system, real-time image processing response and decision control are ensured. At the same time, the exception handling mechanism and fault tolerance control strategy are introduced to improve the stability and reliability of the system in complex environments. In addition, the system also has self-learning and adaptive capabilities, which can continuously optimize algorithm parameters and control strategies through the analysis and learning of historical data and improve the intelligence level of the system.

3. How MASK RCNN works. Mask RCNN is a convolutional neural network developed based on Faster RCNN. In this way, the crack object in the image can be automatically detected and located, and the crack object can be segmented by the template [13]. The network consists of three modules: backbone, regional generation, and functional.

Residual and feature cone networks are used in backbone networks, and ResNet is used as a convolutional neural network to extract high-level visual features [14]. FPN is combined with the ResNet network, and the method of down sampling and up sampling is used to achieve the effective fusion of low-resolution, robust semantic features and high-resolution weak features.

3.1. Supervised Mask RCNN. In the case of guidance, Mask RCNN needs to preprocess a training picture. Each picture is annotated with a picture label system, and the crack location is marked. Then, a

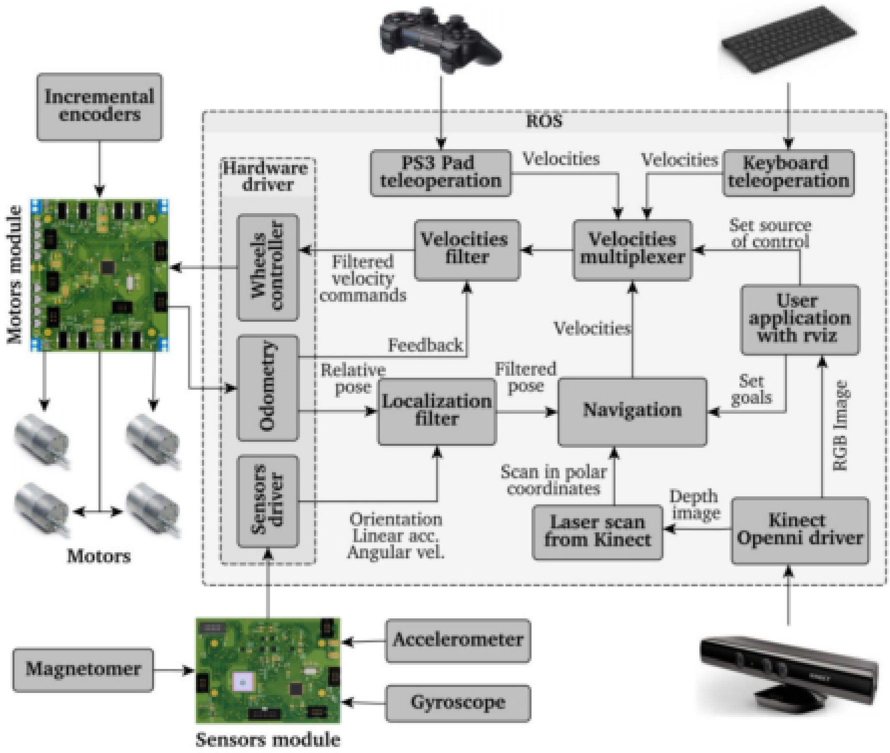


Fig. 2.3: Structure diagram of robot control system.

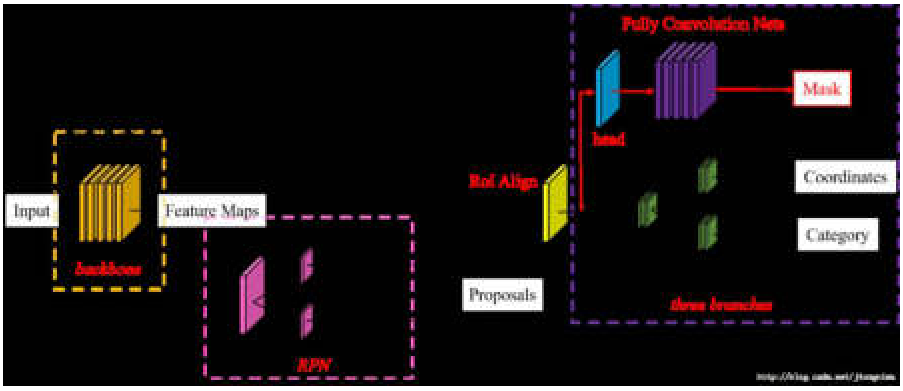


Fig. 3.1: Mask RCNN structure.

rotation method based on gray level, saturation, contrast, and other parameters is proposed to enhance the learning performance of the learning model and avoid overfitting [15]. The crack identification model was established by adjusting the learning rate, weight decay and iteration times. The high-resolution UAV image is input into the crack identification model for accurate location.

3.2. Loss function of Mask RCNN. This paper presents a learning method based on the BP neural network. Then, the loss of the network is minimized by learning the neural network's learning. Loss is the penalty for inaccurate prediction during neural network training [16]. The loss function is used to estimate the

Table 4.1: Analysis table of experimental results.

Speed (cm/s)	Picking target	Experimental picking times	scene images /n	target identification pictures /n	completed picks/n	Identify the extraction rate /%
	Mandarin orange	10	326	320	10	99.09
1	Strawberries	10	318	313	10	99.39
	Mandarin orange	10	260	249	10	97.07
2	Strawberries	10	251	242	10	97.78
	Mandarin orange	10	204	190	9	94.04
3	Strawberries	10	200	187	10	94.34

learning cost of the network.

$$H = H_z + H_1 + H_n \tag{3.1}$$

H is the total loss of the network. $_{\underline{U}}$ is classification loss, which measures the accuracy of the network classification. H_1 is a function of the estimated error, which measures the accuracy of the boundary. H_n is mask loss, which is used to measure the positioning accuracy of the mask [17]. For each class of v, find H_z with the logarithm of the loss function of SoftMax

$$H_z(q, v) = -\log_2 (q_v) \tag{3.2}$$

Find H_1 by the loss function smooth H_1

$$H_1 \left(i^v, \gamma\right) = \sum_{i=\{x,y,w,h\}} \text{smooth}_{H_1} \left(i_i^v - \gamma_i\right) \tag{3.3}$$

$q = (q_0, \cdots, q_j)$ is a value calculated for the function SoftMax. $\gamma = (\gamma_x, \gamma_y, \gamma_w, \gamma_h)$ are the coordinates of the actual edge box of the measured object. $i^v = \left(i_x^v, i_y^v, i_w^v, i_h^v\right)$ is the coordinate correction for the border frame of class v. The loss function of smooth$_H$, looks like this

$$\text{smooth}_{H_1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| < 0.5 & |x| \geq 1 \end{cases} \tag{3.4}$$

H_n is very similar to H_{z1}. It is calculated using the mutual entropy loss function of the two means.

4. Experimental results and comparative analysis.

4.1. Experimental results. Combined with multi-layer RCNN network architecture, the collected results are tested and analyzed. The faster the arm of the acquisition robot moves, the lower the accuracy of feature extraction [18]. The recognition rates of citrus and strawberry are 99.09% and 99.39%, respectively, when the robot arm moves at a low speed of 1 cm/s. The results showed that the practical components of citrus and strawberry were 97.07% and 97.78%, respectively, at 2 cm per second. At 3 cm per second, the effective recognition of citrus and strawberries reached 94.04% and 94.34%, respectively. The analysis of specific experimental results is shown in Table 4.1.

4.2. Comparative experimental analysis. The Mask RCNN model based on Mask RCNN and Faster RCNN were compared to test the performance of the MASKRCNN model on the harvesting robot arm. In this project, two different neural network architectures are used to establish the corresponding learning models for strawberry harvesting, and the corresponding learning models are applied to the harvesting robot arm for the experiment. By collecting image information, the image information of Mask RCNN and Faster RCNN in various harvest stages is studied [19]. The training effect of Mask RCNN is better than Faster RCNNs when

Table 4.2: Comparison of experimental results.

Neural network framework	Speed (cm/s)	Picking target	Experimental picking times	scene images /n	target identification pictures /n	completed picks /n	Identify the extraction rate /%
	1		10	318	313	10	99.39
Mask RCNN	2	Strawberries	10	251	242	10	97.78
	3		10	200	187	10	94.34
	1		10	318	287	10	91.11
Faster RCNN	2	Strawberries	10	251	224	10	90.40
	3		10	200	178	9	89.80

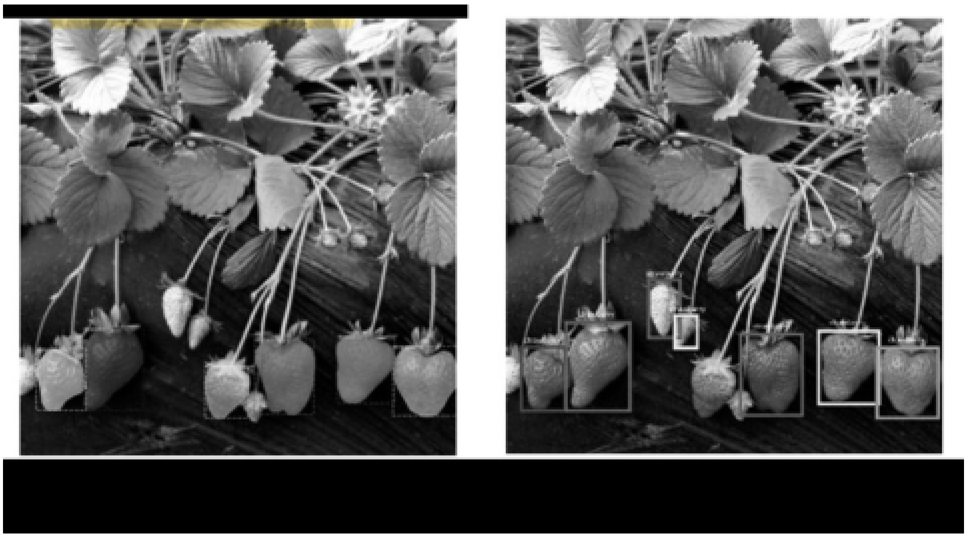


Fig. 4.1: Comparison of recognition of Mask RCNN and Faster RCNN.

the training rate is increased. Faster RCNN can't tell when a strawberry is ripe, as Mask RCNN can. The results of the comparative tests are shown in Table 4.2.

It can divide the categories of pixel level to improve the recognition ability of objects. Because the multi-layer neural network dramatically influences the object identification process, the multi-layer neural network can quickly classify the object in the collection process. Figure 4.1 is a diagram of Mask RCNN compared to Fast RCNN.

5. Conclusion. This research successfully designed and tested an intelligent orchard-picking system for agricultural machinery based on image-processing technology. The system integrates binocular vision technology and the Mask RCNN algorithm, effectively solving the complex problem of fruit accurate recognition and positioning. Through image processing technology, the system can efficiently analyze the orchard environment image and realize real-time detection and recognition of fruit. The binocular vision technology provides depth information for the system and dramatically improves the accuracy of fruit positioning. The Mask RCNN algorithm ensures the high accuracy of fruit recognition in complex backgrounds. This research focuses on the collaborative optimization of hardware selection and software algorithms to ensure the system's stable operation in the actual orchard environment. The simulation results show that the system can accurately identify and locate the fruit in the simulated environment, which verifies the effectiveness of the system design. This research promotes the application of image-processing technology in intelligent agricultural machinery and provides strong support for developing intelligent orchards in the future. Future work will focus on further optimizing the algorithm and expanding the system's capabilities to deal with more diverse orchard environments and more complex picking tasks.

REFERENCES

[1] Zhou, H., Wang, X., Au, W., Kang, H., & Chen, C. (2022). Intelligent robots for fruit harvesting: Recent developments and future challenges. Precision Agriculture, 23(5), 1856-1907.

[2] Zhang, C., Valente, J., Kooistra, L., Guo, L., & Wang, W. (2021). Orchard management with small unmanned aerial vehicles: A survey of sensing and analysis approaches. Precision agriculture, 22(6), 2007-2052.

[3] Tang, Y., Qiu, J., Zhang, Y., Wu, D., Cao, Y., Zhao, K., & Zhu, L. (2023). Optimization strategies of fruit detection to overcome the challenge of unstructured background in field orchard environment: A review. Precision Agriculture, 24(4), 1183-1219.

[4] Dewi, T., Mulya, Z., Risma, P., & Oktarina, Y. (2021). BLOB analysis of an automatic vision guided system for a fruit picking and placing robot. International Journal of Computational Vision and Robotics, 11(3), 315-327.

[5] Shaikh, T. A., Mir, W. A., Rasool, T., & Sofi, S. (2022). Machine learning for smart agriculture and precision farming: towards making the fields talk. Archives of Computational Methods in Engineering, 29(7), 4557-4597.

[6] Boatwright, H., Zhu, H., Clark, A., & Schnabel, G. (2020). Evaluation of the intelligent sprayer system in peach production. Plant Disease, 104(12), 3207-3212.

[7] Bai, Y., Mao, S., Zhou, J., & Zhang, B. (2023). Clustered tomato detection and picking point location using machine learning-aided image analysis for automatic robotic harvesting. Precision Agriculture, 24(2), 727-743.

[8] Javaid, M., Haleem, A., Khan, I. H., & Suman, R. (2023). Understanding the potential applications of Artificial Intelligence in Agriculture Sector. Advanced Agrochem, 2(1), 15-30.

[9] Dewi, T., Risma, P., & Oktarina, Y. (2020). Fruit sorting robot based on color and size for an agricultural product packaging system. Bulletin of Electrical Engineering and Informatics, 9(4), 1438-1445.

[10] Hu, X., Sun, L., Zhou, Y., & Ruan, J. (2020). Review of operational management in intelligent agriculture based on the Internet of Things. Frontiers of Engineering Management, 7(3), 309-322.

[11] Ma, Y., Zhang, W., Qureshi, W. S., Gao, C., Zhang, C., & Li, W. (2021). Autonomous navigation for a wolfberry picking robot using visual cues and fuzzy control. Information Processing in Agriculture, 8(1), 15-26.

[12] Davidson, J., Bhusal, S., Mo, C., Karkee, M., & Zhang, Q. (2020). Robotic manipulation for specialty crop harvesting: A review of manipulator and end-effector technologies. Global Journal of Agricultural and Allied Sciences, 2(1), 25-41.

[13] Chen, L., Wallhead, M., Reding, M., Horst, L., & Zhu, H. (2020). Control of insect pests and diseases in an Ohio fruit farm with a laser-guided intelligent sprayer. HortTechnology, 30(2), 168-175.

[14] Sharma, S., Verma, K., & Hardaha, P. (2023). Implementation of artificial intelligence in agriculture. Journal of Computational and Cognitive Engineering, 2(2), 155-162.

[15] Sun, Y., Ding, W., Shu, L., Li, K., Zhang, Y., Zhou, Z., & Han, G. (2021). On enabling mobile crowd sensing for data collection in smart agriculture: a vision. IEEE Systems Journal, 16(1), 132-143.

[16] Xu, R., & Li, C. (2022). A modular agricultural robotic system (MARS) for precision farming: Concept and implementation. Journal of Field Robotics, 39(4), 387-409.

[17] Chen, Q., Li, L., Chong, C., & Wang, X. (2022). AI-enhanced soil management and smart farming. Soil Use and Management, 38(1), 7-13.

[18] Beloev, I., Kinaneva, D., Georgiev, G., Hristov, G., & Zahariev, P. (2021). Artificial intelligence-driven autonomous robot for precision agriculture. Acta Technologica Agriculturae, 24(1), 48-54.

[19] Boursianis, A. D., Papadopoulou, M. S., Gotsis, A., Wan, S., Sarigiannidis, P., Nikolaidis, S., & Goudos, S. K. (2020). Smart irrigation system for precision agriculture—The AREThOU5A IoT platform. IEEE Sensors Journal, 21(16), 17539-17547.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 22, 2024
Accepted: Aug 28, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1496–1503, DOI 10.12694/scpe.v26i3.4328

MULTI-TARGET VITAL SIGN DETECTION BY FUSION OF BIOLOGICAL RADAR AND CONVOLUTIONAL NEURAL NETWORK

HONGBIN YUAN[*], CHENYAO YUAN[†], AND HUIQUN CAO [‡]

Abstract. In order to address the increasing demand for vital sign detection, the author proposes a multi-target vital sign detection research that combines biological radar and convolutional neural network. Based on the fundamental architecture of convolutional neural networks (CNNs), the author combines classification-based CNN object detection techniques to develop a biological radar multi-target vital sign detection platform. The feasibility of this approach is confirmed through experiments, demonstrating the integration of biological radar and CNNs for multi-target vital sign detection. The experimental results indicate that the biological radar achieves a recognition accuracy of 96.1%, proving the effectiveness of the biological radar detection algorithm. The research on multi-target vital sign detection based on the fusion of biological radar and convolutional neural network is an effective auxiliary method that can provide reference for relevant researchers.

Key words: Biological radar, Convolutional neural network, Multi objective, Vital sign detection

1. Introduction.

In recent years, with the in-depth research and popularization of artificial intelligence algorithms represented by convolutional neural networks, intelligent electronic devices and related application scenarios have become ubiquitous, such as object detection, facial recognition, smart healthcare, etc. CNN achieves high accuracy in model prediction at the cost of high computational complexity by increasing the network structure. Some application scenarios that require high real-time performance not only demand high accuracy of the network, but also require high processing speed of the network [1]. Human vital signs primarily consist of physiological parameters such as heart rate, respiratory rate, body temperature, and blood pressure. These metrics are essential for assessing an individual's health status. Heart rate indicates the number of heartbeats per minute, whereas respiratory rate refers to the number of breaths taken per minute. For healthy adults, the normal respiratory rate ranges from 12 to 20 breaths per minute, and the typical heart rate ranges from 60 to 100 beats per minute. The commonly used detection methods for life signals nowadays include electrocardiography (ECG), photoplethysmography (PPG), and other detection methods that require direct contact with the human body. However, in some special situations such as burns, infectious diseases, and psychiatric patients, the use is restricted. To address emerging needs, research on non-contact vital sign detection technology has garnered significant attention. This technology primarily monitors vital signs using methods such as infrared, electromagnetic waves, and video [2,3].

Non contact vital sign detection does not require physical contact with the target being tested, and can detect distant targets. It not only avoids the constraints of cumbersome wiring harnesses and electrodes, expands its application range, but also avoids psychological pressure on the target during the testing process, making the detection results more realistic. It has a wide range of applications in home and medical health monitoring, driver life status monitoring, and other fields. As an emerging physiological signal detection method, microwave biological radar can detect vital sign signals such as human heart and lung activity. Compared with traditional methods such as electrocardiogram and pulse, microwave biological radar technology is not only non-contact, but also has good penetrability, which can penetrate obstacles such as clothing and bedding for detection. These advantages make microwave biological radar technology have potential applications in medical diagnosis, health monitoring, disaster rescue and other fields. This study aims to combine biological radar and CNN technology

[*] School of Telecommunications and Intelligent Manufacturing, Sias University, Zhengzhou, Henan, 451150, China. (Corresponding author, kfyhb@163.com)

[†] School of Computer and Software Engineering, Sias University, Zhengzhou, 451150, China.

[‡] Academic Affairs Office of Kaifeng Modern Technology Secondary Vocational School, Kaifeng, Henan, 475000, China.

to achieve accurate detection and analysis of multi-target vital signs, providing new technological solutions for fields such as medical diagnosis and health monitoring [4].

2. Literature Review. With the continuous development of medical technology, there is an increasing demand for the application of vital sign signal monitoring in scenarios such as infant monitoring and smart elderly care. Traditional vital sign detection methods such as electrocardiogram monitors, smart bracelets, oximeters, etc. require contact with the patient's body, which is not only inconvenient but also prone to cross infection. Biological radar is a new concept radar technology that mainly focuses on the human body as the detection object. It can detect signals such as breathing, heartbeat, and body movement of the human body in a non-contact manner, and has become a research hotspot at home and abroad in recent years. Compared with traditional contact detection, biological radar can reduce the discomfort and psychological burden caused by electrodes, wires, sensors, etc. on the human body, and has broad application prospects in the field of social medicine. Liu, J. et al. introduced a convolutional neural network leveraging multi-scale feature fusion to enhance the integration quality of multimodal medical images. The findings show that this method outperforms other state-of-the-art techniques across most metrics [5]. Zhang, A. et al. developed a multimodal fusion convolutional neural network (MFCNN) that employs a dual-stream convolutional neural network (CNN). This network extracts shared information from surface electromyography (sEMG) and accelerometer signals from various subjects [6]. Mohan, R. et al. designed a convolutional neural network called MIDNet18, a tailored medical image analysis and detection network, to diagnose various lung diseases using chest CT images. The MIDNet-18 CNN architecture has simplified model construction, minimal complexity, simple techniques, and high-performance accuracy, and can classify binary and multi class medical images [7].

To tackle these challenges and establish a foundation for multi-target vital sign detection, the author suggests a study focused on detecting multiple vital signs through the integration of biological radar and convolutional neural networks. Non-contact vital sign detection technology offers a valuable supplementary method for health monitoring. Developing a device for detecting human vital signs using microwave radar technology. Design a continuous wave radar circuit based on the Doppler principle to detect human micro motion signals; Using short-time Fourier transform and interpolation algorithm to extract human heart rate and respiratory rate parameters; Introducing embedded platforms to achieve miniaturization design and integration of devices; Develop embedded signal processing software to achieve real-time signal processing, recording, and display. With the growth of training data and advancements in machine performance, convolutional neural network-based object detection has surpassed the limitations of traditional methods, becoming the leading algorithm in contemporary object detection.

3. Research Methods.

3.1. Convolutional Neural Networks.

3.1.1. Basic Structure of Convolutional Neural Networks. Convolutional neural networks were proposed by Professor LeCun at the University of Toronto in Canada. The earliest convolutional neural networks were used as classifiers, mainly for image recognition [8]. In Figure 3.1, a convolutional neural network (CNN) is depicted as a hierarchical model comprising essential components: an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer. CNNs are specialized for image processing, leveraging weight-sharing through convolution to extract features in the convolutional layers. This architecture enables the network to progressively extract hierarchical features from low-level to high-level representations. These high-level features are then classified using fully connected and output layers, producing one-dimensional vectors that categorize the input image. Thus, CNNs can be conceptually divided into a feature extractor (input, convolutional, and pooling layers) and a classifier (fully connected and output layers), each contributing distinct functions in the image recognition process.

3.1.2. Classification based Convolutional Neural Network Object Detection. Traditional object detection methods involve a sequence of steps: preprocessing, window sliding, feature extraction, feature selection, feature classification, and post-processing. In contrast, convolutional neural networks (CNNs) encompass the capabilities of feature extraction, selection, and classification within their architecture. This integrated approach enables CNNs to streamline object detection tasks by reducing the need for separate preprocessing steps

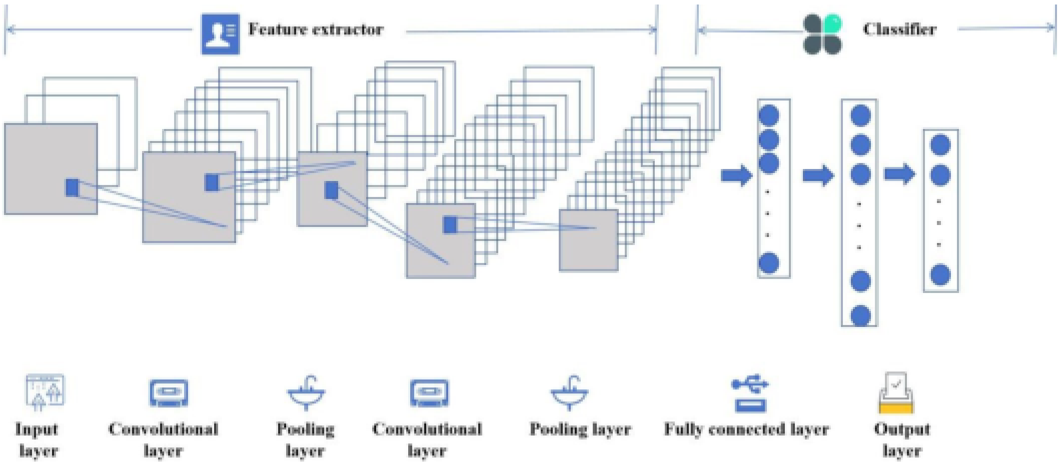


Fig. 3.1: Basic Structure of Convolutional Neural Network

and directly processing raw data to classify objects effectively [9]. Convolutional neural networks (CNNs) are capable of directly classifying candidate regions generated by sliding windows, a method known as classification-based CNN object detection. Unlike traditional object detection methods with multiple steps, this approach simplifies the process to three main steps: sliding windows, image classification, and post-processing, where sliding windows and post-processing methods are predefined. As a result, research in this area predominantly centers on enhancing CNNs' abilities in feature extraction, selection, and classification to improve the overall accuracy of image recognition.

Researchers have developed algorithms to extract sub-images containing specific semantic information from the target image, thereby reducing the number of candidate regions. These regions vary in size and represent distinct semantic meanings. By employing convolutional neural networks (CNNs) for classification and recognition, this approach enables the detection of objects across multiple scales and classes. This method significantly enhances the efficiency of object detection by focusing computational efforts on relevant and meaningful regions within the image [10]. In the evolution of object detection methods, researchers are exploring new approaches to enhance accuracy by reconfiguring convolutional neural networks (CNNs) as regressors. Instead of relying solely on classification, these methods treat the entire image as a potential candidate region. This involves directly inputting the image into the CNN to regress and pinpoint the precise position information of the target within the image itself. This approach represents a shift towards more holistic and integrated methodologies in CNN-based object detection [11].

3.1.3. SIMD Computing Acceleration Method for Convolutional Networks Based on ARM Processor. ARMv8 has 32 128 bit registers, and the author uses embedded assembly programming based on ARMv8 to implement the design of SIMD convolution accelerator. One instruction is used to perform multiplication and accumulation operations on 8 sets of data [12]. Table 3.1 lists several common instructions, namely addition, subtraction, multiplication, multiply accumulate, data load, and store instructions.

3.2. Bioradar Multi target Vital Signs Detection Platform.

3.2.1. Composition and Detection Principle of Experimental Platform. Figure 3.2 shows the overall structure of the experimental platform, which mainly includes the biological radar sensor, radar baseband signal conditioning circuit, MSP430F 5529LaunchPad experimental board, and upper computer data processing system. Through the experimental platform, non-contact respiratory and heartbeat signals can be collected, processed, and analyzed. The experimental steps are as follows: One is that the biological radar sensor emits high-frequency continuous microwaves to the human body; The second is to use the radar baseband signal conditioning circuit to perform DC offset correction, filtering, and amplification processing on the signal output by the radar sensor, filter out noise in the signal, and preserve the effective frequency components of the

Table 3.1: Several Common SIMD Instructions

op	Output operands	Input operands
add	v0.8h,	v0.8h,v8.8h
sub	v0.8h,	v0.8h,v8.8h
mul	v0.8h,	v0.8h,v8.8h
mla	v0.8h,	v0.8h,v8.8h
1d4	{v0.8h,v1.8h,v2.8h,v3.8h}	[%0]
st4	[%2]	{v0.8h,v1.8h,v2.8h,v3.8h}

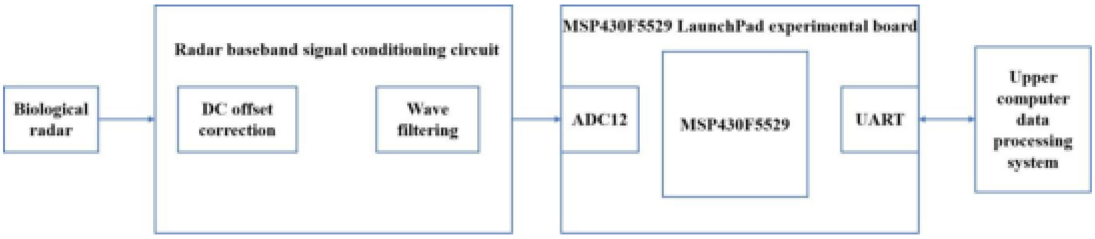

Biological radar
Radar baseband signal conditioning circuit
DC offset correction
Wave filtering
MSP430F5529 LaunchPad experimental board
ADC12
MSP430F5529
UART
Upper computer data processing system

Fig. 3.2: Overall structure of experimental platform

body motion signal; The third is to use the MSP430F5529 LaunchPad experimental board to collect and process the body motion signals output by the conditioning circuit, and transmit the processing results to the upper computer data processing system through a serial communication interface. The upper computer data processing system, often developed using MATLAB or LabVIEW, serves to process and analyze collected body motion signals. Its primary function includes extracting respiratory and heartbeat signals from the gathered data.

The front-end antenna of the continuous wave biological radar emits high-frequency continuous microwaves to the human body [13]. When the microwave irradiates the human body, the chest wall micro motion caused by respiratory movement and heart beat modulates the reflected echo signal in phase and frequency. After signal amplification, filtering and other processing, respiratory, heartbeat and other signals are extracted from the echo signal.

Microwave biological radar provides a non-contact and penetrating means of detecting biological motion information [14]. Among them, continuous wave radar uses the Doppler effect of electromagnetic waves to detect the displacement, velocity and other motion information of targets, with a simple structure and easy processing of received signals. Continuous wave biological radar emits continuous electromagnetic waves towards human targets, while receiving echoes reflected from the human body surface. By changing the frequency or phase of the echo signal, micro motion information on the body surface is extracted and calculated [15]. Due to the higher frequency of electromagnetic waves, the stronger the reflection at the interface between human skin and air, but at the same time, the reflection on obstacles such as clothing and bedding will also increase. In order to achieve higher detection accuracy and reduce the power of clutter, biological radar usually uses a carrier frequency of (2.4-60) GHz. In terms of physiology, the micro motion information on the human body surface can reflect certain physiological activities of the human body, such as detecting chest wall vibrations to obtain heart and lung activity related information such as breathing and heartbeat. The amplitude of surface mechanical vibration caused by normal human heartbeat movement is about 0.6 mm; The amplitude generated by respiration is around (4-12) mm. If a 10GHz frequency band biological radar is used to detect chest wall motion, every 1mm displacement of the chest wall will cause a maximum phase shift of 25.2. Therefore, theoretically, although the amplitude of chest wall vibration is small, the phase shift reflected in the radar baseband can still be distinguished when the carrier frequency is high enough [16].

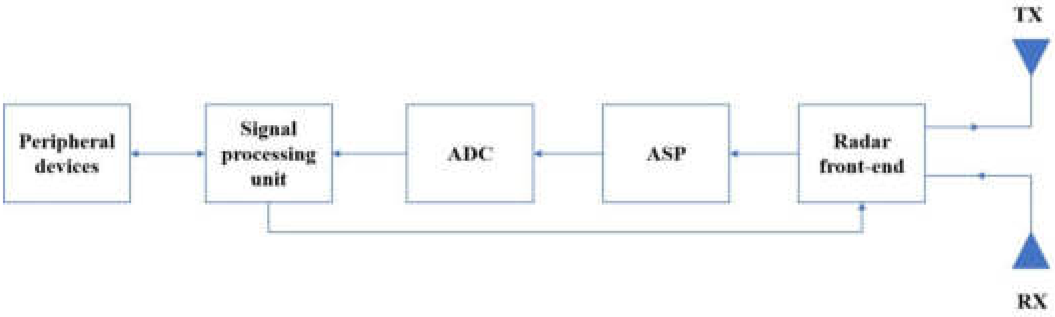


Fig. 3.3: System Principle Block Diagram

The sine signal generated by the oscillator is represented as the radar transmission signal by formula 3.1:

$$T(t) = A_T cos(2\pi f t + \phi(t)) \tag{3.1}$$

The distance traveled by the transmitted signal to become the received signal is represented by formula 3.2:

$$R(t) = A_R cos[2\pi f t - \frac{4\pi d_0}{\lambda} = \frac{4\pi x(t - d(t)/c)}{\lambda} + \phi(t - \frac{2d_0}{c} - \frac{2x(t - d(t)/c)}{c} + \theta_0)] \tag{3.2}$$

The received signal is mixed with the signal in the radar receiver and low-pass filtered to obtain the baseband signal output, which is expressed as formula 3.3:

$$B(t) \approx A_B cos[(\frac{4\pi d_0}{\lambda} - \theta_0) + \phi(t) - \phi(t - \frac{2d_0}{c}) + \frac{4\pi x(t)}{\lambda}] \approx A_B cos[\varphi(d_0) + \frac{4\pi x(t)}{\lambda}] \tag{3.3}$$

3.2.2. System Hardware Design. When extracting the target's cardiopulmonary activity information from the received signal of the biological radar, the useful signal frequency is between (0.2 10) Hz and the amplitude is weak. The front-end circuit of the continuous wave radar requires a sufficiently high signal-to-noise ratio, and performs DC offset correction, signal amplification, analog bandpass filtering, and analog-to-digital conversion on the down converted received signal. Then, various physiological parameter detection and extraction algorithms are used to obtain the required information. The system diagram of the vital sign detection device is shown in Figure 3.3.

The radar front-end uses equal amplitude sine wave transmission, zero intermediate frequency receiver structure, and the radar carrier operates in the frequency band of 10 GHz; The antenna used for transmission and reception is a microstrip antenna to save space; Analog signal processing circuit (ASP) is used to amplify, filter and level shift signals. The analog filter uses a lower cutoff frequency of 0.1 Hz to suppress DC offset and low-frequency noise, and an upper cutoff frequency of 100 Hz to prevent signal sampling aliasing; The analog-to-digital conversion uses a 16 bit high-precision ADC; The signal processing unit is an embedded platform used for digital signal processing and regulating the work of various interconnected units; At the same time, the signal processing unit is connected to various peripheral devices to achieve the output of vital sign results, including display, alarm, data storage, and communication with computers; On the other hand, it realizes user control signal input, including functions such as controlling the operation of the system, setting various system parameters, etc. [17].

3.3. Experimental research.

3.3.1. Detection experiment using analog signal source. The individual differences in human cardiovascular activity and the resulting surface vibrations vary with changes in body condition, posture, and environment. Therefore, in order to conduct quantitative, controllable, and reproducible experimental research, the author first used a device that simulates human chest wall vibration as the detection target [18].

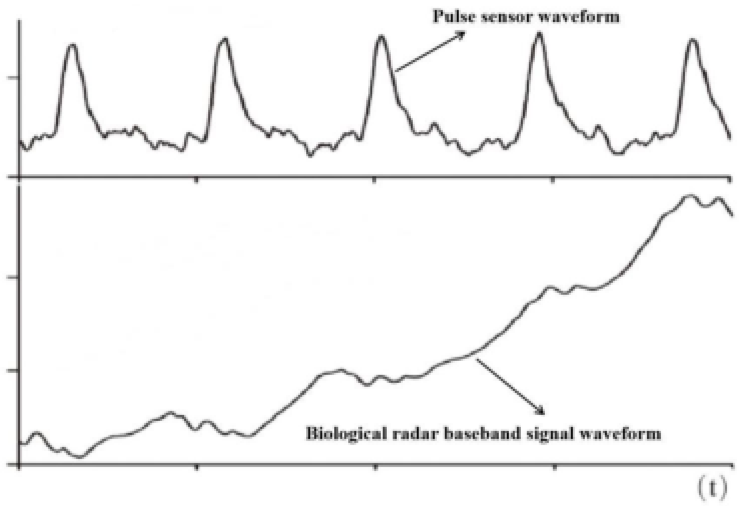


Fig. 3.4: Pulse sensor waveform and biological radar baseband signal waveform

The device reads in typical mechanical vibration waveform files generated by the computer on the surface of the human chest wall, converts them into vibration signals, and outputs them to simulate human cardiopulmonary activity. The vibration surface of the analog signal source device is covered with a metal layer to obtain strong electromagnetic wave reflection ability.

Referring to the physiological signal surface vibration parameters mentioned above, configure the analog signal source to send a sine wave with a vibration amplitude of 4 mm and a frequency of 1 Hz, and place it 50 cm away from the biological radar to obtain the baseband signal spectrum of the receiver. Both noise and useful signal energy are concentrated in the low frequency range, with a baseband signal amplitude of 10.64 dBmV at 1 Hz. Due to the nonlinearity of the demodulation system, harmonics appear in the spectrum. The system base noise is -94.64 dBmV, and the baseband signal signal-to-noise ratio is 98.36 dB.

For the detection of vital signs of real human targets, experimental results vary with factors such as the actual environment and subject status. The author fixed the distance between a single stationary human target and the biological radar at 30 cm, and used a pulse sensor to synchronously collect the pulse signal of the human target as a reference signal for heart rate recognition results, for comparative analysis. Figure 3.4 shows the comparison between the waveform of the pulse sensor and the waveform of the biological radar baseband signal, where the heartbeat signal of the biological radar baseband signal is superimposed with the larger amplitude respiratory signal. As shown in the figure, the biological radar simultaneously detects respiratory and heartbeat related information of human targets, where the heartbeat signal corresponds well with the reference signal; In addition, when using biological radar for the analysis and extraction of heartbeat signals, due to the strong baseline drift interference provided by respiratory signals, it is not easy to directly analyze using conventional heart rate measurement methods such as peak seeking or zero crossing detection in the time domain.

4. Results Analysis. In order to evaluate the accuracy of heart rate recognition using the aforementioned vital sign detection algorithm, an experiment was conducted to collect a signal containing changes in the target human heart rate, as shown in Figure 5, to obtain the heart rate recognition results of the biological radar and reference signal after passing through the detection algorithm. When the difference between the results obtained by the biological radar and the reference signal is less than 2%, it is considered that the biological radar recognition is correct. The calculation shows that for the data in Figure 4.1, the recognition accuracy of the biological radar is 96.1%. It can be seen from this that the effectiveness of the biological radar detection algorithm, in addition, the detection algorithm can quickly track and smoothly transition to changes in heart rate. Through experiments, it has been found that the biological radar detection system can effectively detect the respiratory and heart rate of the target in real time within a range of 90 cm from the target. When the

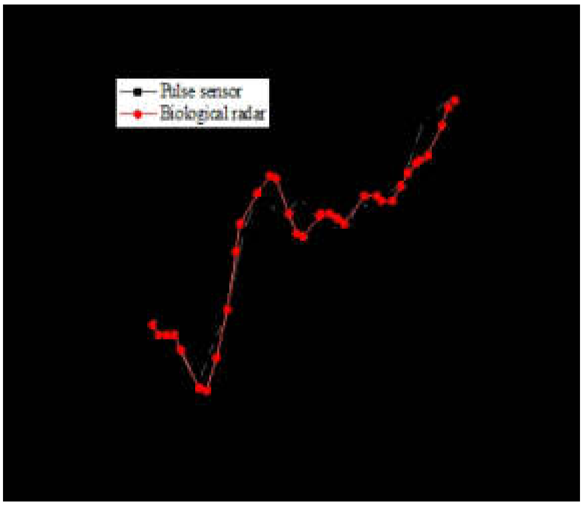


Fig. 4.1: Heart rate extraction results

distance is farther, due to the influence of noise in the actual environment, the noise power will be equivalent to the peak power of the heartbeat, causing fluctuations in the heart rate detection results and affecting the accuracy of heart rate recognition.

5. Conclusion. The current biological radar detection device needs to consider how to automatically adjust the phase shift constant in hardware, so that the demodulation point is near the optimal demodulation point to improve detection accuracy; In terms of algorithms, the tracking algorithm used can eliminate sudden strong interference signals but lacks sufficient suppression of continuous clutter interference, resulting in a decrease in the accuracy of actual human detection and recognition. Further research should not be limited to finding the highest spectral peak, but should attempt to use other in-depth spectral analysis algorithms. In order to address the increasing demand for vital sign detection, the author proposes a multi-target vital sign detection research that combines biological radar and convolutional neural network. The effectiveness of the method is verified through experiments.

Acknowledgement Fund project: Henan Province Science and Technology Research Project (232102210195)

REFERENCES

[1] Kavitha, K., & Banu, D. S. (2024). Genetic algorithm framework for 3d discrete wavelet transform based hyperspectral image classification. Journal of the Indian Society of Remote Sensing, 52(3), 645-657.

[2] Park, J. H., & Jung, S. (2023). Form-finding to fabrication: a parametric shell structure fabricated using an industrial robotic arm with a hot-wire end-effector. Nexus network journal: Architecture and mathematics, 25(4), 829-848.

[3] Chen, C., Zhang, T., Teng, Y., Yu, Y., Shu, X., & Zhang, L., et al. (2022). Automated segmentation of craniopharyngioma on mr images using u-net-based deep convolutional neural network. European Radiology, 33(4), 2665-2675.

[4] Wu, C., Liao, X., & Zhou, W. M. (2023). Full-reference image quality assessment via low-level and high-level feature fusion. International journal of pattern recognition and artificial intelligence, 37(11), 1.1-1.20.

[5] Liu, J., Zhang, L., Guo, A., Gao, Y., & Zheng, Y. (2023). Multi-scale feature fusion convolutional neural network for multimodal medical image fusion. Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things, 42(1), 199-200.

[6] Zhang, A., Li, Q., Li, Z., & Li, J. (2023). Multimodal fusion convolutional neural network based on semg and accelerometer signals for intersubject upper limb movement classification. IEEE sensors journal(11), 23.

[7] Mohan, R., Rama, A., & Ganapathy, K. (2022). Comparison of convolutional neural network for classifying lung diseases from chest ct images. International Journal of Pattern Recognition and Artificial Intelligence, 9(2), A_213-A_222.

[8] Bendory, T., Lan, T. Y., Marshall, N. F., Rukshin, I., & Singer, A. (2023). Multi-target detection with rotations. Inverse Problems and Imaging, 17(2), 362-380.

[9] Borisov, V. V., & Misnik, A. E. (2023). Ontological engineering in complex systems based on meta-associative graphs. Pattern recognition and image analysis: advances in mathematical theory and applications in the USSR, 33(3), 234-241.

[10] Rubanga, D. P., Cuevas, S. A. M., Arvelyna, Y., & Shimada, S. (2022). Detection of geographic faults using deep learning model from dem and remote sensing data in djibouti. Journal of Arid Land Studies, 32(3), 88-88.

[11] Xiao, R., Zhang, Y., Wang, B., Xu, Y., Fan, J., & Shen, H., et al. (2023). A low-power in-memory multiplication and accumulation array with modified radix-4 input and canonical signed digit weights. IEEE transactions on very large scale integration (VLSI) systems, 31(11), 1700-1712.

[12] Pan, H., Zou, Y., & Gu, M. (2022). A spectrum estimation approach for accurate heartbeat detection using doppler radar based on combination of ftpr and twv. EURASIP Journal on Advances in Signal Processing, 2022(1), 1-22.

[13] Zhang, N., Zeng, J., Lv, P., Miao, X., Chen, C., & Lin, J. (2023). Comparison between a modified fast 3-dimensional turbo spin-echo and diffusion-weighted imaging with background suppression in evaluation of lumbosacral plexus and its branches. Journal of Computer Assisted Tomography, 48(1), 156-160.

[14] Kyong-Hee, N., & Ro, L. J. (2023). Safety management regulation and practice standards on living modified organism (lmo) facilities under the ministry of environment. Plant biotechnology reports, 17(6), 787-802.

[15] Maxwell, M., Tooley, T., & Penvose I.Gehrke C.Koueiter D.Wiater B.Baker E.Wiater J.M. (2023). Evaluating trunnionosis in modular anatomic shoulder arthroplasties: a retrieval study. Journal of shoulder and elbow surgery, 32(10), 1999-2007.

[16] Kadam, V., Deshmukh, A., & Bhosale, S. (2023). Hybrid beamforming for dual functioning multi-input multi-output radar using dimension reduced-baseband piecewise successive approximation. International Journal of Engineering (IJE), 36(1), 182-190.

[17] Suzuki, M. (2022). Data-driven tuning based on virtual internal model tuning considering input limitation. Transactions of the Society of Instrument and Control Engineers, 58(10), 491-493.

[18] Varga, C. M., Kwiatkowski, K. J., Pedro, M. J., Groepenhoff, H., Rose, E. A., & Gray, C., et al. (2022). Observation of aerosol generation by human subjects during cardiopulmonary exercise testing using a high-powered laser technique: a pilot project. Journal of Medical and Biological Engineering, 42(1), 1-10.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 22, 2024
Accepted: Aug 27, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1504–1510, DOI 10.12694/scpe.v26i3.4329

ATHLETES' PHYSICAL FITNESS EVALUATION MODEL BASED ON DATA MINING

DACHENG GU*

Abstract. This study firstly introduces the design and implementation of a physical health monitoring bracelet and describes in detail how the bracelet collects multi-dimensional physiological data such as heart rate, step number and sleep quality of athletes. This paper then discusses the application of data mining algorithms in processing these data, including cluster analysis, classification and prediction models, to identify the key influencing factors of athletes' physical fitness. This paper uses the RFID anti-collision algorithm to improve the accuracy and efficiency of data acquisition, avoiding the error and delay that may occur in traditional methods. The practicability and validity of the model are verified by system simulation. The simulation results show that the model can accurately evaluate the athletes' physical fitness and provide real-time feedback and personalized training suggestions for the coaching team. This not only helps athletes improve their competitive performance but also prevents sports injuries during daily training.

Key words: Athletes' physical health; Physical health monitoring bracelet; Data mining; RFID anti-collision algorithm; System simulation

1. Introduction. With the rapid development of information technology, wearable devices have gradually integrated into People's Daily life. Especially in sports, they have provided unprecedented convenience for athletes' training and competition. In recent years, the physical health of athletes has become one of the hot spots of sports research, and wearable bracelets, as a convenient data collection tool, have played an essential role in this aspect. Previous studies have explored various ways to assess athletes' physical fitness. For example, literature [1] proposes an athlete fatigue evaluation method based on heart rate variability, which determines the fatigue state of athletes by analyzing their heart rate data, thus providing a basis for training adjustment for coaches. However, this method mainly focuses on a single heart rate index and fails to reflect the athlete's physical condition fully. With the progress of science and technology, physical health monitoring bracelets came into being. These bracelets can monitor several physiological parameters of athletes in real time, such as heart rate, blood oxygen saturation, sleep quality and so on. Literature [2] introduces the design and implementation of a multi-functional physical health monitoring bracelet, which integrates various sensors and can monitor the physiological indicators of athletes in real time. It transmits it via Bluetooth to a mobile device for analysis. The appearance of this bracelet provides convenience for the real-time monitoring of athletes' physical health. However, relying solely on the wristband to collect data is not enough. How to effectively process and analyze these data is the key. This requires the use of data mining technology. Data mining is a process of extracting useful information from large amounts of data, which can help us discover the patterns and trends behind the data. In athletes' physical health, data mining technology has many application prospects. For example, literature [3] used data mining technology to analyze the training data of athletes and found the key factors affecting athletes' performance, providing valuable reference information for coaches. Accuracy and completeness are critical when dealing with large amounts of data. Some scholars have introduced the RFID anti-collision algorithm. RFID is a wireless communication technology that can identify a specific target through radio waves and read the relevant data. The integration of RFID technology in the physical health monitoring bracelet can realize the simultaneous reading and data exchange of multiple bracelets, thus avoiding the problem of data conflict and loss. Literature [4] proposes an anti-collision algorithm based on RFID technology, which can effectively solve the collision problem in the process of multi-tag identification and improve the accuracy and efficiency of data reading. In addition to the above literature, many scholars have

*Department of Physical Education and Health, Nanning Normal University, Nanning, Guangxi, 530001, China (Corresponding author, dcgu2024@163.com)

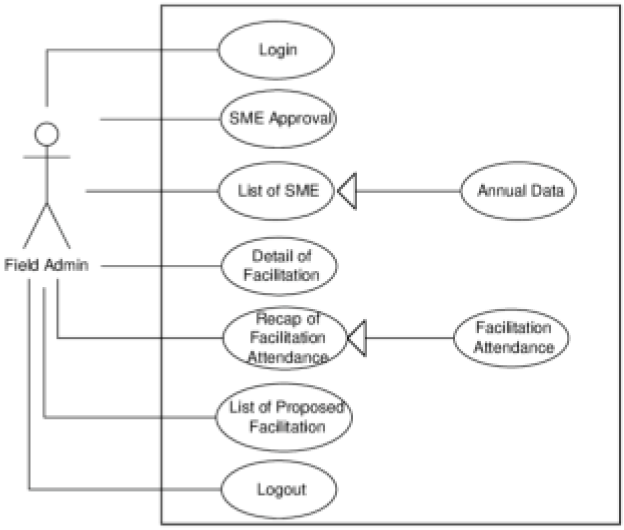


Fig. 2.1: Super Administrator function diagram.

conducted in-depth research on athletes' physical health, physical health monitoring bracelets, data mining and RFID anti-collision algorithms. For example, literature [5] discussed the relationship between athletes' physical health and sports injuries and put forward targeted preventive measures; Literature [6] studied the application of data mining technology in athletes' mental health assessment; Literature [7] gives a comprehensive review on the application of RFID technology in the field of sports.

This paper will start with the actual needs of athletes' physical health, combined with wearable bracelets, data mining and RFID anti-collision algorithms and other advanced technologies, to build a complete set of athletes' physical fitness data mining evaluation models [8]. The model will realize real-time monitoring, scientific evaluation, and personalized guidance of athletes' physical conditions, positively contributing to sports development in China.

2. Demand analysis and model architecture. The information network management model of physical education quality monitoring in colleges and universities is proposed based on data and network support. This paper introduces the computer-aided instruction system in the computer-aided instruction system. It includes test comparison, dynamic student interaction information collection and other functional modules [9]. The construction, release, implementation and feedback of the results of the system are introduced into the specific management mode. This provides a solid basis for formulating school sports plans and related policies.

2.1. Requirement Analysis.

2.1.1. Super Administrator. The super administrator has the most rights and is responsible for managing and assigning work to all administrators. Usually, the developer of the system is personally responsible. As shown in Figure 2.1, the super administrator's job is relatively easy, just adding, adjusting, and delegating tasks to lower administrators.

2.1.2. System Administrator. Since there are many job types of system administrators, there should be more people with system administrator rights. The system administrator is responsible for daily system maintenance, bug repair, student data proofreading, data backup, student file management, examination and statistics [10]. It can be manually repaired when an exception occurs or recovered if the student's login information is lost or forgotten.

2.1.3. Student users. This study takes students as the primary research object. It includes system login, test booking, personal information filling and modification, examination score inquiry, feedback verification, information submission and other modules.

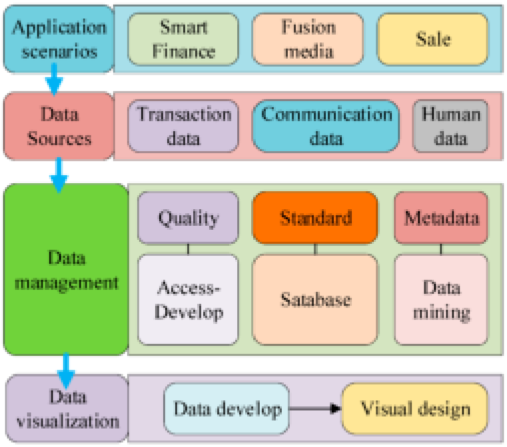


Fig. 2.2: Detailed architecture of the system.

2.2. System Architecture. The system adds a front end of physical fitness detection and acquisition devices. The three modules cooperate to complete the collection and storage of the module. The front end of the physical fitness testing device takes the measuring device as the main body to realize the field collection of students' physical fitness [11]. When the data is returned to the server, it is based on the specifications for body recognition [12]. The corresponding logical parsing is performed on the server, and then the parsing results are stored in the system's database. Publish information through the web in a unified standard. Its construction relationship is shown in Figure 2.2.

2.3. System hardware design. The bus control unit is the system's center, and other functional units are connected to the bus control unit. Each control unit works together to complete the functions required by the system. The MAX30100 microcontroller is used to detect human health. The LIS3DH three-way accelerometer is used as the communication module to obtain the acceleration. The communication module has 13.56Mhz passive RF electronic marking [13]. This machine adopts a 2.2V lithium-ion battery and a BQ24040 charger that can be charged with 5V. The screen uses organic light-emitting diodes to bend the screen. The Buzzer module adopts an active buzzer with an alarm function. P0.26 and P0.27 in the primary control component circuit U1 are connected to the XC1 pin, and the XC2 pin is connected to the 16/32 MHZ crystal oscillator to generate oscillation. U2 module, U4 module, U5 module, U8 module is the physical monitoring module, communication module, display module, buzzer module: power modules U6 and U7 supply power to the control module U1 through a voltage stabilization circuit. The pins SPC, SDI, SDO, and CS arranged by the mobile component U3 are successively connected with the pins P0.10, P0.11, P0.12, and P0.13 of the main control component UI so that information can be exchanged between the two components [14]. The communication component U4 is configured with different pins; pin 2 uses RF pins, and the output impedance is 50 ohms. No.1 and No. 11 are digital power supply pins that can be directly connected to the 3.3V working power supply, 27/28 is connected to the crystal oscillator circuit, which can stimulate the vibration frequency, 25,26,32 is connected to the central control module pins P0.14, P0.152, P0.16. This line realizes the data communication between the student's wristband and the teacher's terminal. This paper provides a contactless intelligent RFID identification tag, S50, which can improve the security of module operation and effectively resist external disturbance. The type S50 electronic marking comprises an RF interface and a digital controller [15]. Its memory is 1 KB. Each section has 16 blocks, and each block is 16 bytes. The buzzer module uses a Tereski-type integrated circuit connected to the central control component through the input and output pins. Under the adjustment of the central control unit through the low-level switch, the S8550 triode drives the internal oscillation source to achieve the purpose of alarm.

3. RFID anti-collision algorithm. Given the shortcomings of the existing monitoring methods, such as poor stability and poor real-time performance, this project intends to introduce the RFID collision avoidance method of equal partition to achieve optimal control of the human body and embed it into the smartwatch to improve its practicality [16]. Divide A picture into A segment with B as the number of marks to be identified. The matching probability is:

$$F(R=r)=Z_n^r\cdot\left(\frac{1}{\tau}\right)^r\cdot\left(1-\frac{1}{\tau}\right)^{n-r} \tag{3.1}$$

r is an integer and its value interval is $r\in[0,n]$. If it's $r=1$, then you get the odds of a match. The calculation formula for setting the initial time slot value $\delta_h^{\tau,n}, \delta_s^{\tau,n}$ and conflicting truth value $\delta_e^{\tau,n}$ is as follows:

$$\delta_h^{\tau,n}=\tau\cdot F_h=n\cdot\left(1-\frac{1}{\tau}\right)^{n-1} \tag{3.2}$$

$$\delta_s^{\tau,n}=\tau\cdot F_s=\tau\cdot\left(1-\frac{1}{\tau}\right)^{n} \tag{3.3}$$

$$\delta_\varepsilon^{\tau,n}=\tau\cdot F_e=\tau-\delta_h^{\tau,n}-\delta_s^{\tau,n} \tag{3.4}$$

Set the throughput rate of the system to D_{RFID} , and the ratio of the marks recognized by a frame identifier to the total number of time slots is:

$$D_{RFID}=\frac{\delta_h^{\tau,n}}{\tau}=\frac{n}{\tau}\cdot\left(1-\frac{1}{\tau}\right)^{n-1} \tag{3.5}$$

Figure 3.1 describes in detail the specific implementation process of the anticollision algorithm of the RFID system (the picture is quoted in Thinking about the Strategy and Practice Path of Modern Agricultural Industry Development in the context of Big Data).

If the number of marks is more than 354, grouping the marks into N groups by equal area is necessary. The signal is randomly selected in a specific time slot. The number of multiple labels was predicted by the DFSAC-II method. Optimize the time slot based on accurate data. When the tag is identified, the following tag identification process starts after the "+1" group, and the method does not end until all the tags have been tagged.

4. Experimental results and analysis.

4.1. Checking System Performance. The value range of the number of markers is set to [50,1500], and the initial value of the method is set to 256. Set the value of the method to 1500. The resulting average is the result [17]. The total number of time slots of the proposed algorithm, FSA_256 and DFSA algorithms change with increasing the number of tags (fig.4.1). The delay of the proposed algorithm increases with the increase in the number of tags. When the number of tags is 4912, it is reduced by 83.23% and 84.21% compared with FSA_256 and DFSA, respectively, which proves that the algorithm can effectively shorten the collision time and save the system identification time [18]. The change trend of system throughput is shown in Figure 4.2. The algorithm's delay increases with the increase in the number of tags. When the number of tags is 4912, it decreases by 83.12% and 84.32% compared with FSA_256 and DFSA, respectively, which proves that the algorithm can effectively reduce the conflict slot and save the system identification time [19]. The results show that the network traffic changes under various methods with increased markers.

4.2. Physical monitoring module software debugging. The memory of the MAX30100 is accessed using the IIC interface. First, it is initialized. When the red light is on, it is serial debugging so that it can be measured in real-time. The data is sent to the PC through the serial port and compared with the physiological indicators obtained by the red light and infrared LED. The MAX30100 transmits the collected information

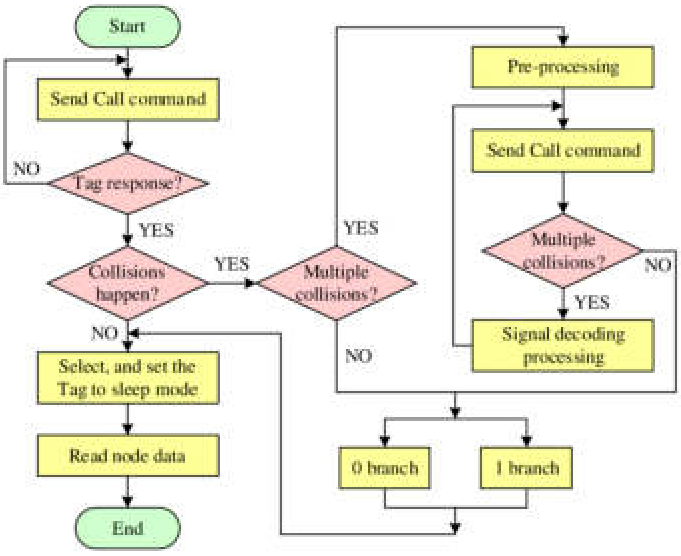


Fig. 3.1: Flow of RFID anti-collision algorithm.

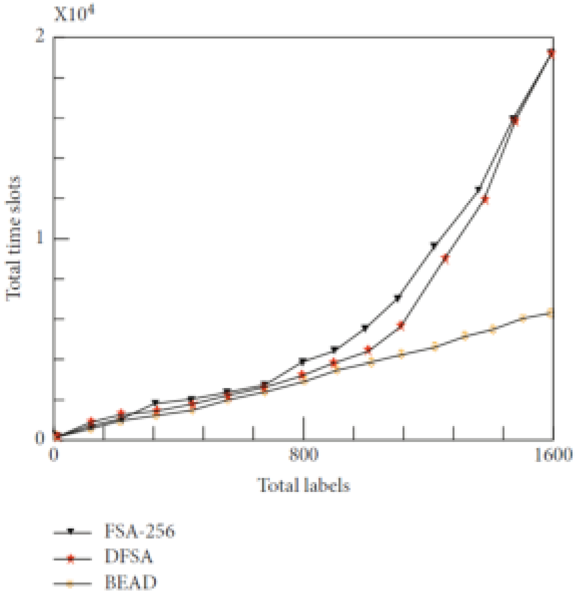


Fig. 4.1: Comparison results of total time slots of the proposed algorithm, FSA_256 and DFSA algorithms.

to the master controller and stores it in memory. The electrical signal is converted into parameters such as blood oxygen saturation and heart rate through calculations. The developed physical fitness monitoring module can accurately display the athlete's heartbeat, blood oxygen concentration and other indicators to reflect the athlete's physical fitness better. Through the motion control of the hand ring, the monitored motion information can be accurately displayed on the OLED screen, which further verifies that the design of the motion module is correct.

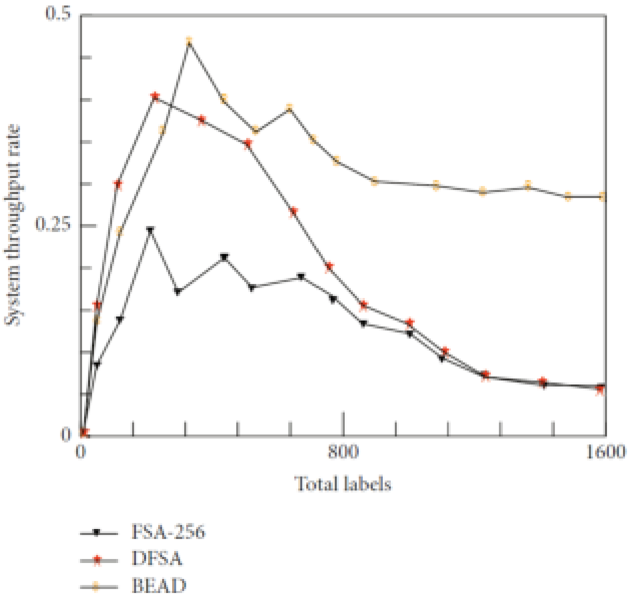


Fig. 4.2: Comparison of total time slots.

5. Conclusion. This study successfully constructed a data mining evaluation model of athletes' physical fitness based on wearable bracelets. Through the in-depth analysis of a large number of physiological data collected by physical health monitoring bracelet, we find that data mining technology has great application potential in this field. In particular, the combination of RFID anti-collision algorithms improves the accuracy and efficiency of data acquisition and enhances the model's robustness and practicability. The establishment of this model provides a new technical means for real-time monitoring and scientific evaluation of athletes' physical health. It can accurately identify the key indicators of athletes' physical fitness and provide scientific and comprehensive training guidance for the coaching team. At the same time, the model can also prevent sports injuries in daily training and improve athletes' competitive performance.

REFERENCES

[1] Yuliandra, R., & Fahrizqi, E. B. (2020). Development of endurance with the ball exercise model in basketball games. Jp. Jok (Jurnal Pendidikan Jasmani, Olahraga Dan Kesehatan), 4(1), 61-72.

[2] Kalkhoven, J. T., Watsford, M. L., & Impellizzeri, F. M. (2020). A conceptual model and detailed framework for stress-related, strain-related, and overuse athletic injury. Journal of science and medicine in sport, 23(8), 726-734.

[3] Stoyel, H., Slee, A., Meyer, C., & Serpell, L. (2020). Systematic review of risk factors for eating psychopathology in athletes: a critique of an etiological model. European Eating Disorders Review, 28(1), 3-25.

[4] Rozali, M. Z., Puteh, S., Yunus, F. A. N., Hamdan, N. H., & Latif, F. M. (2022). Reliability and validity of instrument on academic enhancement support for student-athlete using Rasch Measurement Model. Asian Journal of University Education, 18(1), 290-299.

[5] Lee, J., & Zhang, X. L. (2021). Physiological determinants of VO2max and the methods to evaluate it: A critical review. Science & Sports, 36(4), 259-271.

[6] Ramirez-Campillo, R., García-Hermoso, A., Moran, J., Chaabene, H., Negra, Y., & Scanlan, A. T. (2022). The effects of plyometric jump training on physical fitness attributes in basketball players: A meta-analysis. Journal of sport and health science, 11(6), 656-670.

[7] Li, B., & Xu, X. (2021). Application of artificial intelligence in basketball sport. Journal of Education, Health and Sport, 11(7), 54-67.

[8] Yuliandra, R., Nugroho, R. A., & Gumantan, A. (2020). The Effect of Circuit Training Method on Leg Muscle Explosive Power. Active: Journal of Physical Education, Sport, Health and Recreation, 9(3), 157-161.

[9] Malikov, N., Konoh, A., Korobeynikov, G., Korobeynikova, L. E. S. I. A., Dudnyk, O., & Ivaschenko, E. (2020). Physical condition improvement in elite volleyball players. Journal of Physical Education and Sport, 20(5), 2686-2694.

[10] Paquette, M. R., Napier, C., Willy, R. W., & Stellingwerff, T. (2020). Moving beyond weekly ”distance”: optimizing quantification of training load in runners. journal of orthopaedic & sports physical therapy, 50(10), 564-569.

[11] Mossman, L. H., Slemp, G. R., Lewis, K. J., Colla, R. H., & O'Halloran, P. (2024). Autonomy support in sport and exercise settings: A systematic review and meta-analysis. International Review of Sport and Exercise Psychology, 17(1), 540-563.

[12] Kim, M., Do Kim, Y., & Lee, H. W. (2020). It is time to consider athletes' well-being and performance satisfaction: The roles of authentic leadership and psychological capital. Sport Management Review, 23(5), 964-977.

[13] Upadhyay, A. K., & Khandelwal, K. (2022). Metaverse: the future of immersive training. Strategic HR Review, 21(3), 83-86.

[14] Ekstrand, J., Spreco, A., Windt, J., & Khan, K. M. (2020). Are elite soccer teams' preseason training sessions associated with fewer in-season injuries? A 15-year analysis from the Union of European Football Associations (UEFA) elite club injury study. The American journal of sports medicine, 48(3), 723-729.

[15] Towlson, C., Salter, J., Ade, J. D., Enright, K., Harper, L. D., Page, R. M., & Malone, J. J. (2021). Maturity-associated considerations for training load, injury risk, and physical performance in youth soccer: One size does not fit all. Journal of sport and health science, 10(4), 403-412.

[16] Broglio, S. P., McAllister, T., Katz, B. P., LaPradd, M., Zhou, W., & McCrea, M. A. (2022). The natural history of sport-related concussion in collegiate athletes: findings from the NCAA-DoD CARE Consortium. Sports medicine, 52(2), 403-415.

[17] Luczak, T., Burch, R., Lewis, E., Chander, H., & Ball, J. (2020). State-of-the-art review of athletic wearable technology: What 113 strength and conditioning coaches and athletic trainers from the USA said about technology in sports. International Journal of Sports Science & Coaching, 15(1), 26-40.

[18] Impellizzeri, F. M., Menaspà, P., Coutts, A. J., Kalkhoven, J., & Menaspà, M. J. (2020). Training load and its role in injury prevention, part I: back to the future. Journal of athletic training, 55(9), 885-892.

[19] Flockhart, M., Nilsson, L. C., Tais, S., Ekblom, B., Apró, W., & Larsen, F. J. (2021). Excessive exercise training causes mitochondrial functional impairment and decreases glucose tolerance in healthy volunteers. Cell metabolism, 33(5), 957-970.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 22, 2024
Accepted: Aug 28, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1511–1521, DOI 10.12694/scpe.v26i3.4335

VIRTUAL REALITY ASSISTED TEACHING SYSTEM FOR IMPROVING ENGLISH READING COMPREHENSION ABILITY

HONGYING YANG[*] AND ZHENQIU YANG[†]

Abstract. In order to enhance students' enthusiasm for learning English, the author designed an English reading assisted teaching system based on virtual reality technology. System hardware design, selecting HTG Vive head mounted devices and high-performance computers to build the architecture; System software design, collecting teaching images through virtual graphics cards, using hierarchical network coding technology to remotely transmit teaching data, building a virtual teaching environment, and using collision detection algorithms to achieve human-computer interaction teaching. The test results indicate that the system has an average response time of 29.14 seconds for different quantities of English reading resources, demonstrating good operational performance. The traditional way of reading English is too monotonous and boring. Virtual reality technology has increased the fun of reading and profoundly influenced the transformation of English reading classrooms.

Key words: Virtual Reality Technology, English reading, Human-computer interaction, Assisted teaching

1. Introduction. Reading is an activity in which people acquire information and understand the world. Through reading, people understand, comprehend, and absorb knowledge, change their thinking, and enhance their cognition. With the development of the Internet in the information age, ”fragmented reading” and ”fast food entertainment” methods such as microblog, circle of friends, and short videos have become the mainstream way of information intake. Traditional reading methods face greater challenges, and it is crucial to seek reading development [1]. Language learning includes four aspects: listening, speaking, reading, and writing, among which ”reading” occupies an important position, especially for English learners. English reading not only plays a crucial role in English learning activities, but also plays a key role in improving overall language learning abilities. The value and significance of English reading goes far beyond just increasing knowledge. Reading not only cultivates English learners' sense of language, but also promotes the accumulation of English vocabulary, improves their English reading and writing abilities [2]. From this perspective, English reading is an important means and effective way to learn English well. Meanwhile, the improvement of English reading ability is of great help in enhancing the overall English proficiency of English learners. The key to improving reading proficiency is to master the methods and enhance efficiency. How to combine the skills and methods of English reading with modern technology and English reading, explore new paths and methods of reading, enhance English learners' interest in English reading, and conduct English discourse reading more efficiently and quickly has always been a topic of discussion for English educators and learners [3]. The Ministry of Education has proposed the construction of a national level virtual simulation experimental teaching center. Virtual Reality VR technology, as a multidimensional experience platform, has brought new opportunities and challenges to the development of English reading.

Virtual reality (VR) technology is a computer simulation system used to create and experience virtual worlds. This simulation system combines virtuality with reality and has its own uniqueness. It can use computer systems to generate a virtual simulation environment, allowing participants to immerse themselves in this simulation environment and provide them with an immersive experience. This immersive experience can mobilize multiple senses to create a sense of immersive participation for users. Virtual reality technology can transform a single text content in a book into a relatively multidimensional and three-dimensional reading environment, allowing users to experience an immersive reading experience. This immersive reading experience

[*] Department of Foreign Languages, Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei, 066000, China.

[†] Department of Foreign Languages, Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei, 066000, China (Corresponding author, yzq8008@126.com)

is different from traditional regular reading, as it allows readers to fully immerse themselves in this multidimensional and three-dimensional reading environment [4]. This three-dimensional, rich, and multi angle reading method enhances readers' interest in reading and also improves reading efficiency. For English educators, this multidimensional reading experience can make the classroom more diverse, allowing students to experience the situation more vividly, enhancing the interactivity of the classroom, and thus improving teaching efficiency. By applying virtual reality technology to the classroom, teachers can better stimulate students' creative thinking and innovative abilities, allowing them not only to understand the content in the text, but also to connect with the future through these contents, and generate valuable ideas for the future through creative thinking. This is the deeper meaning of reading. The application of virtual reality technology in the education industry can enhance students' interest in learning and strengthen their ability to master English reading by creating a teaching environment of ”self-directed learning and human-computer interaction”. Therefore, a virtual reality technology-based English reading teaching system is designed. This system breaks the traditional teaching concept by transforming students' passive learning into active learning, and transforming modular English reading into dynamic virtual scenes; Students improve their English reading ability by engaging in conversations with virtual characters; Human computer interaction allows students to ask questions in real-time in the teaching system, which then provides feedback to students or teachers, achieving diversified teaching methods.

2. Literature Review. With the rapid advancement of global technology, coupled with the rapid development of communication and network technology in China in recent years, China's VR technology has also made significant progress under this trend. Based on the characteristics of VR technology and actual project requirements, science and engineering subjects are the areas that best reflect the advantages of VR technology. Especially in disciplines such as architecture, mechanics, and physics. With the use of advanced technology, learners have a new way of learning. In virtual laboratories, learners can intuitively observe various well-known buildings and carefully study and research the various technical details contained in the buildings, thus breaking the limitations of traditional book teaching and graphic display. Virtual reality technology combines multimedia and graphic simulation techniques, allowing learners to fully immerse themselves in a virtual environment. With the popularization of computer and VR technology, the way of learning English is shifting from traditional textbooks and classroom teaching to computer-assisted language learning (CALL), and even towards gamified learning. This transformation provides learners with a more immersive experience, stimulating their interest and participation in learning.Yuanxuan, M.A. et al. investigated the factors and regulatory strategies that affect second language reading ability through a scale survey. It has important theoretical and practical value in improving the reading ability of second language learners [5]. Ostovar Namaghi et al. conducted a study on the impact of interactive games on English learners' reading comprehension ability and game attitudes. They found through evaluating participants' perceptions that they hold a positive attitude towards this type of game.The research results indicate that interactive games are an effective means to enhance students' participation, motivation, and learning outcomes in the classroom. In addition, the study provides some practical suggestions for future in-depth exploration [6]. Lijun, H. U.'s research found that in the teaching mode of virtual reality, teachers are no longer the leaders of the reading classroom, but supporters, participants, and researchers in the reading process of students. The reshaping of this role not only helps to strengthen good interaction between teachers and students, but also promotes communication among students, bringing new vitality and motivation to the classroom, thereby improving students' reading level [7]. Su, C. and his co authors delved into the connotation and characteristics of ESA theory from its perspective, and explored the feasibility of applying ESA theory in business English reading teaching. They analyzed the main problems in business English reading teaching in vocational colleges and proposed a series of new reform strategies to evaluate learning effectiveness and improve teaching level [8]. Liu, X. studied the implementation steps of comprehensive English literacy teaching and established a digital comprehensive English literacy teaching model. Research shows that combining English literacy with digital teaching directly affects students' overall English learning outcomes [9].

With the maturity of virtual reality technology and the widespread use of high-performance hardware, the opportunity to integrate virtual reality technology into English reading comprehension skills is also becoming more mature. By transforming it into an intuitive sensory experience, learners will definitely be deeply impressed by the entire process, and their corresponding memories will also be more profound. In addition, due to the

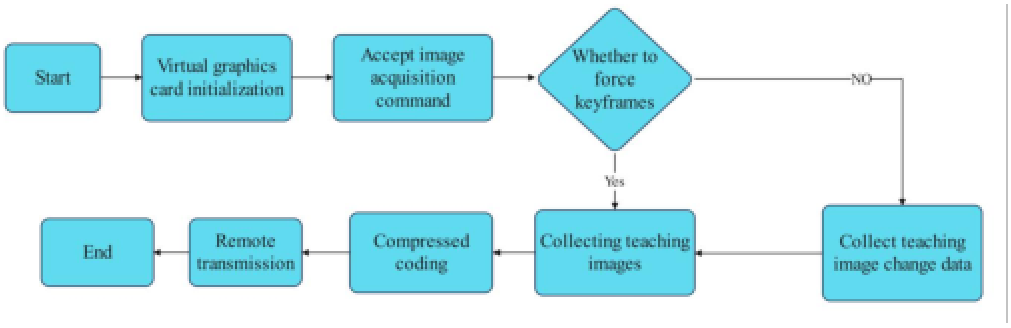

Start
Virtual graphics card initialization
Accept image acquisition command
Whether to force keyframes
NO
Yes
Collect teaching image change data
Collecting teaching images
Compressed coding
Remote transmission
End

Fig. 3.1: Flow Chart for Collecting Teaching Images

ability of VR videos to stimulate learners' creativity, applying VR videos to daily English reading teaching may produce unpredictable 'chemical reactions'.

3. Design of English Reading Assistance Teaching System.

3.1. Hardware Design. Applying virtual reality technology to English reading can greatly improve the enthusiasm of English learners and bring good reading results. The hardware architecture of the English reading assisted teaching system mainly consists of computers, virtual reality (VR) devices, displays, and audio equipment. Select a high-performance computer with 8GB of RAM and an Intel i54 CPU as the data layer to process 3D models, audio, and image files for English reading instruction. There are many types of VR devices on the market, among which the HTC Vive headset has a large number of application development interfaces, making it very suitable for the English reading assisted teaching system designed by the author. The HTC Vive device is mainly composed of a headset, a positioning device, and a joystick, among them, two positioning devices are equipped with laser positioning sensors, so they can accurately locate moving targets without relying on cameras, track moving targets wearing head mounted devices and control handles, and follow the positioning targets within a certain range. To apply virtual reality technology in the education system, it is necessary to establish three-dimensional models of teaching content and enhance students' immersion through virtual models. In the process of modeling English teaching, traditional hardware devices cannot simulate such models, and virtual digital models are generated through computers.

3.2. Software Design.

3.2.1. Collecting Teaching Images. The main function of the auxiliary teaching system is to assist teachers in teaching. When teachers use the system to teach, to ensure the quality of teaching, the system must accurately capture real-time images and audio data of the teacher, transmit them to the server for processing, and then display the processed content to students within virtual scenes [10]. By capturing image changes through a virtual graphics card, the system avoids the need for full-screen copying and processing, thus significantly reducing the workload involved in processing screen data. Figure 3.1 illustrates the program flow for capturing teaching images.

As depicted in Figure 3.1, when teachers employ the author's system for remote teaching, they begin by initializing the virtual graphics card and awaiting the image acquisition command from the teaching system. Upon receiving this command, the teacher commences remote teaching. During this process, the teaching system automatically captures teaching images. Each image captured undergoes automatic background compression and encoding, after which the compressed image is remotely transmitted to the server for additional processing. By repeatedly running this process, continuous collection of teaching images for teachers can be achieved. The fundamental significance of teaching design is to find out and solve the various problems existing in the actual teaching process.

3.2.2. Remote transmission of teaching data. In order to ensure the stability of remote transmission of teaching data and enable students to learn English reading courses in a stable virtual teaching environment, hierarchical network coding technology is adopted to optimize the remote transmission process of teaching data, obtain the average delay of coding, and ensure a stable transmission port for teaching data [11]. Assuming a directed acyclic graph is used to represent a single source multicast network, where the directed acyclic graph is a set of nodes and directed edges. All nodes in the directed acyclic graph are input with data characters to make it a linear channel set of a finite field. Then, the output channel of the nodes is encoded and processed, and the data characters that need to be forwarded are:

$$Z(h) = \sum_{h' \in ln(h)} C_{h',h} Z(h') \tag{3.1}$$

In the formula, where $Z(h)$ represents the data forwarding character,h denotes the channel arc before forwarding,h' indicates the forwarded channel arc, ln(h) represents the set of input channels for data nodes, and C denotes the secondary data of the Galois field. First, calculate the average delay of network encoding and integrate it into the subnet system.The connection point is the source point of the subnet system and also the destination point of the English reading auxiliary teaching system, ensuring stable remote transmission of teaching data.

3.2.3. Building a Virtual Teaching Environment. In order to make the virtual teaching environment consistent with the actual scene, the size of individual geometries is first adjusted through an error function, and the precise individual geometries are assembled, spliced, and other operations to form the entire virtual teaching environment. Next, the parameters of the virtual teaching environment are adjusted according to the specific requirements of the English reading assistance teaching system. This optimization aims to enhance the display performance of the teaching system by achieving higher resolution in virtual scenes. Therefore, the calculation formula for pixels in virtual teaching environments is

$$\begin{cases} S_x = L_x cos\alpha t N \cdot S_z \\ S_y = L_y cos\alpha t N \cdot S_z \end{cases} \tag{3.2}$$

In the formula, equation 3.2 is utilized to construct a high-resolution virtual teaching environment where students can fully engage in English reading immersion. Here, S_x represents the pixel point in the virtual environment's horizontal direction, S_y in the vertical direction, and S_z the total pixel count. Additionally, L_x and L_y denote the horizontal and vertical dimensions of geometric objects in the virtual setting, while α signifies the lighting angle parameter in real scenes. N stands for the number of surfaces on a single geometric object within the virtual scene. This setup aims to enhance students' interest in English reading through immersive learning experiences in the virtual environment.

3.2.4. Implementing human-computer interaction teaching. In order to determine whether students have made contact with objects in the virtual teaching environment, collision detection algorithms are used to detect whether human-computer interaction has occurred. The collision detection algorithm creates a bounding box outside each geometric object in the virtual teaching environment, and uses the intersection test between the student and the bounding box to determine whether a collision has occurred. Therefore, the expression for axis aligned bounding boxes is 3.3:

$$0 = \{(x_r, y_r, z_r)|x_{min} \leqslant x_r \leqslant x_{max}, y_{min} \leqslant y_r \leqslant y_{max}, z_{min} \leqslant z_r \leqslant z_{max}\} \tag{3.3}$$

In the formula: in the virtual teaching environment, (x_r, y_r, z_r) represents the spatial coordinate of a geometric object, while $x_{min}, y_{min}, z_{min}$ and $x_{max}, y_{max}, z_{max}$ denote the minimum and maximum spatial coordinates of its bounding box, respectively.The procedure for detecting collisions between students and geometric objects in a virtual teaching environment, utilizing equation 3.3, comprises the following steps:Initially, identify and establish the spatial coordinates of every geometric bounding box present within the virtual teaching environment. Next, sort the projection list of the geometric center on each coordinate axis; Finally, based on the sorting results, it is determined whether the student coordinates overlap with the bounding box coordinates, in order to achieve the most realistic human-computer interaction teaching.

Before conducting formal teaching research, it is necessary to consider the operating environment or hardware support of VR videos and the actual experimental environment in which learners are located. In this teaching research, VR videos are played based on mobile devices, and the requirements for mobile phone configuration are not high. Therefore, before implementation, only a rough understanding of students' mobile phone situations is needed. And the experimental environment is quite important. Due to the immersive nature of VR videos, learners need sufficient activity space when using VR videos for learning, in order to avoid discomfort. At the same time, it is necessary to eliminate external interference to prevent learners from making cognitive errors, which can lead to cognitive dissonance.

The experimental environment refers to the physical space in which learners use VR videos for learning. In this experiment, a teaching laboratory with sufficient space and bright lighting was selected. Because learners need a certain amount of activity space when using VR video, and also need to eliminate external interference, it is necessary to choose a space that meets the above conditions to promote the smooth progress of the experiment. In addition, the overall environment of the experimental site is bright and spacious, making it an ideal experimental site. In this closed experimental environment, learners have enough space to freely rotate their bodies for multi angle observation, resulting in a stronger sense of immersion. The enclosed environment created by the laboratory provides learners with a sense of security and eliminates external interference factors. Learners can repeatedly use various senses for comprehensive feelings and experiences. On the contrary, if there is insufficient space or improper settings, learners will experience anxiety and unease, and even the virtual environment created by VR devices will make learners feel dizzy and disoriented. Learners will have strong rebellious emotions, which completely reverses the purpose of the experiment. Therefore, this study has sufficient space and moderate lighting to ensure that learners have a stable environment and the experimental results are guaranteed [12].

The user's first impression of VR videos mainly comes from the clarity and playback frame rate or smoothness. In this experiment, as long as the software can be opened smoothly, the hardware requirements for the virtual reality videos produced in this study are not high. Choose teaching experiments with ample activity space and bright lighting. At the same time, ensure that there are no foreign objects on the ground to avoid tripping and other situations. Upload relevant resources to the public cloud drive before the experiment officially begins, accompanied by relevant explanatory documents and videos, to help learners master the basic knowledge and skills.

The application implementation process mainly utilizes VR videos to conduct teaching research, test learners' satisfaction with this teaching experiment, and then compare the two with traditional teaching methods. Analyze and evaluate based on these results. Based on the final evaluations, guidance, and comments provided by the learners, comprehensively analyze the shortcomings in this actual production and provide reference for further optimization and improvement in the future. Throughout the implementation process, learners' experiences and sentiments will be collected via questionnaire surveys and interviews. The questionnaire was initially crafted to encompass personal metrics across three dimensions: sensory perceptions, interactive engagement, and cognitive understanding.Then use interview methods to gain a deeper understanding of learners' experiences and enjoyment of VR videos, in order to conduct comprehensive testing and evaluation of VR videos. If the problem is caused by the real environment, simply making modifications on the design side is useless. So, identifying the root causes of various problems during the design phase has a decisive guiding role in the overall direction of the design.

4. System operation testing. In order to test the performance of the system, the traditional teaching system is compared with the designed system to find out its functional differences.

4.1. Determine testing environment and indicators. After the design of the English reading auxiliary teaching system is completed, it is necessary to test it, promptly discover and handle relevant errors in the system, and ensure the reliability of the system operation. The main purpose of this experiment is to test the functionality of the system, namely black box testing. The relevant equipment parameters for setting up the testing environment are shown in Table 4.1. Based on this environment, conduct experiments to test the response time of the designed system and evaluate its performance based on the test results.

The testing indicators this time are rooted in user experience theory. Marketing expert Bernd Schmidt categorizes consumer experience into five primary experiential factors: senses, cognition, creativity, behavior,

Table 4.1: Equipment Parameters for Building Test Environment

Server side			Client	
WEB server	Application server	Communication	PC	Communication
Inter Pentium G3250,Memory 8 GB, CPU frequency 3.2 GHz	Inter Pentium G2020, 8GB memory, CPU clock speed of 2.9 GHz	Bandwidth 8MHz	Inter Pentium G3250,Memory 8 GB, CPU frequency 3.2 GHz	Bandwidth 4MHz

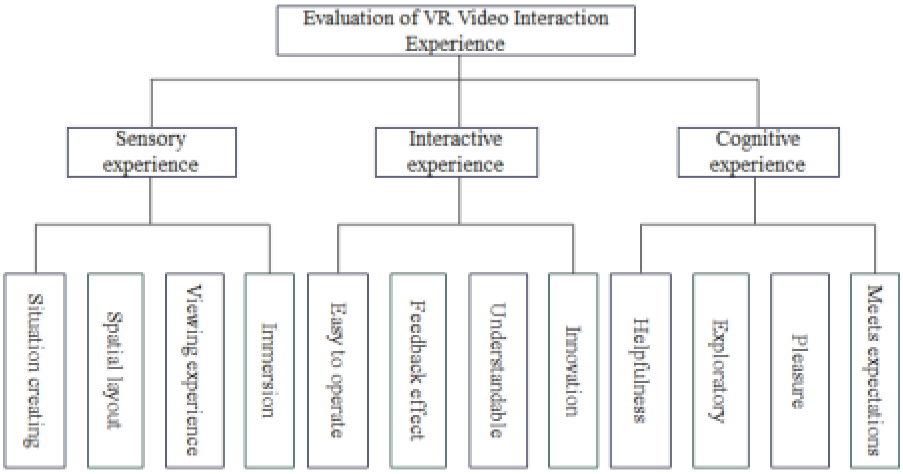


Fig. 4.1: VR Video Evaluation System

and association. VR videos have immersive, interactive, and imaginative qualities. Therefore, the author's implementation also refers to the following aspects of this theory: In evaluating the VR video, we initially focus on sensory immersion, examining whether it provides learners with an enriching sensory experience and a truly immersive environment. Secondly, we assess the overall system's usability, determining if it is intuitive, easy to learn and control, and capable of fostering new cognitive experiences for users. Ultimately, we gauge whether learners are able to utilize the VR video effectively, gaining valuable new content and insights through their learning experience [13]. The evaluation system, depicted in Figures 4.1 and 4.2, is structured into three dimensions according to the aforementioned theory, each containing four specific indicators. A questionnaire has been developed based on these five indicators, where students provide ratings based on the questionnaire prompts. Ratings are conducted on a five-point scale, ranging from 1 to 5, with 1 indicating the lowest satisfaction and 5 indicating the highest satisfaction.

4.2. Analysis of Test Results. A total of 50 people participated in this test, and the final valid answer received was 50 points. Through both horizontal and vertical analysis of the collected data, it is evident that students rated comprehension and cognition highly, scoring 4.3 and 4.5 respectively. However, there is an imbalance in the level of interactive experience, which received a lower score of 3.8. This suggests that while the VR films in this study are effective in providing multisensory experiences, they fall short in terms of interactive engagement.

Among the 12 indicators in Figure 4.3, the item with the highest overall score is the one with clear expression of content. The scores for immersion, overall restoration, and pleasure are also high. However, it performs poorly in terms of interactive experience and feedback experience. This result is actually not surprising. The parts that perform well showcase the unique characteristics of VR videos [14]. The lower scoring part is mainly due to two reasons: Firstly, the overall video production still uses 3D modeling methods, which still have a significant gap between real scene rendering and live action shooting. The second reason is that the production volume is

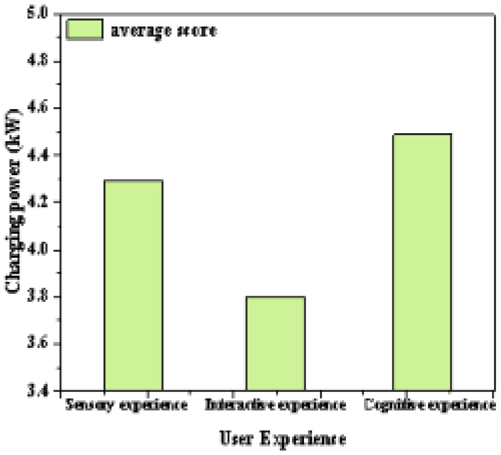


Fig. 4.2: Average Statistics of Three Dimensional User Experience

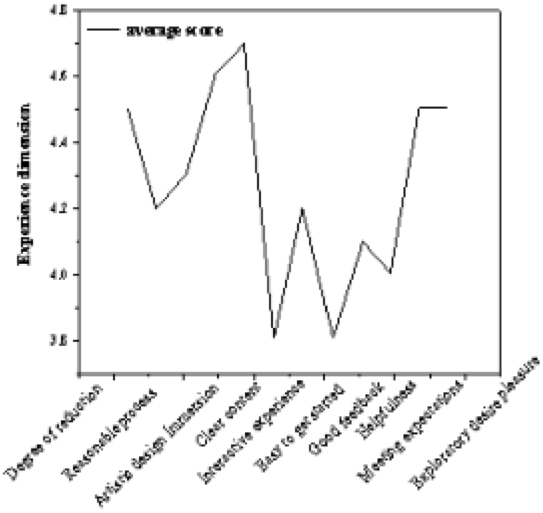


Fig. 4.3: Statistics of Student Satisfaction with VR Videos

large, and it is impossible to take into account all the elements during production.

In this educational research, a comparative survey was conducted to assess learners' experiences with traditional teaching methods versus VR teaching videos. Specific findings are detailed in Table 4.2. The analysis reveals that a majority of students perceive VR teaching videos to more accurately reproduce explained content compared to multimedia materials used in traditional teaching methods.Additionally, VR teaching videos provide a more compelling intuitive experience and immediately capture learners' attention. Moreover, they inherently offer a certain level of enjoyment. However, some students have reported experiencing discomfort

Table 4.2: Comparison between VR Video and Traditional Teaching Methods

Order number	Problem	Traditional teaching methods	VR teaching video
1	The degree of restoration of the explained content		
2	The strength of intuitive perception		
3	The fun of overall design		
4	Do you feel any discomfort during the learning process		
5	Emotional mobilization throughout the entire process		

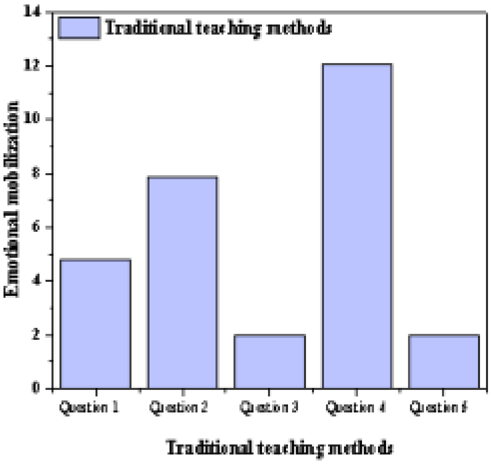


Fig. 4.4: A survey of the attitudes towards traditional teaching methods

while using this technology.Given that some students are new to this technology and may encounter challenges with certain operations and usage methods, these issues can be addressed gradually through future revisions and updates. The final question reveals that most learners believe VR video technology effectively addresses the limitations of traditional textbooks. Converting ordinary content into specific situations and visual representations significantly enhances emotional engagement, specific data are shown in Figures 4.4 and 4.5.

After completing the use of VR teaching videos, interviews were conducted with learners, and some conclusions were drawn by summarizing and analyzing their responses. Ninety percent of individuals are intrigued by this technology, convinced that the intuitive sensations and experiences it offers can create lasting impressions quickly, greatly aiding subsequent learning and memory retention.

In conclusion, the majority of surveyed students maintain a positive outlook on VR video teaching. They expressed a desire for more high-quality VR teaching videos in future educational settings. Based on the findings of this experiment, it's evident that VR teaching videos have considerable potential for growth in educational applications. As a cutting-edge technology with distinct advantages, it has the potential to address some of the inherent challenges in traditional teaching methods [15].

In this system testing experiment, a certain teaching system was selected as the experimental comparison to set different numbers of English reading resources. Two systems were loaded separately to obtain images. The response time results of the two systems to different numbers of English reading resources are shown in Figure 4.6. As shown in Figure 4.6, the English reading assistance teaching system based on virtual reality

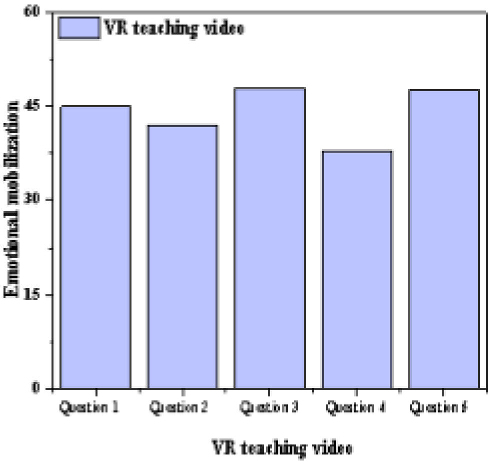


Fig. 4.5: Attitude survey on VR teaching videos

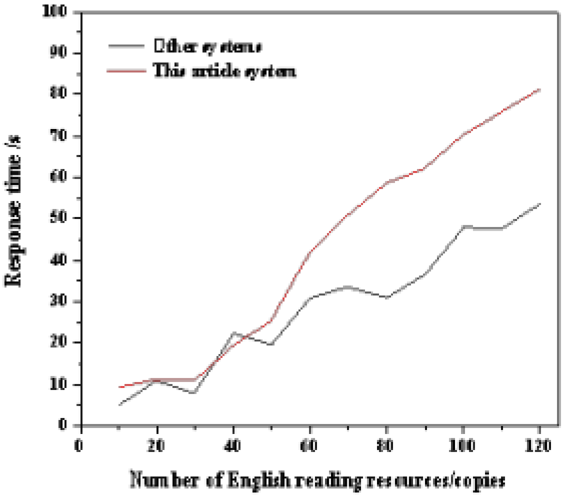


Fig. 4.6: Comparison Curve of System Response Time

technology has a faster response time, better performance, and is more suitable for English reading assistance teaching.

4.3. Overall Effect Evaluation.

4.3.1. Overall Design Evaluation. The design of VR videos highlights key knowledge points and emphasizes commentary, allowing students to better grasp the key knowledge points. From the final result, the overall experience is still quite good. At the same time, the opinions and suggestions raised by students regarding VR videos have been uniformly organized, and finally summarized as follows:

Due to hardware and production limitations, some detailed scenes and actions cannot be well restored.

If more resources can be obtained, better modeling and sound effects can be achieved, further enhancing the learning experience for learners.

There are some shortcomings in the design of interaction. The overall focus is still on explanation and analysis. You can try adding some small tests or interactive elements in real time during the explanation process.

5. Conclusion. In China, English learning remains a significant challenge due to limited exposure to immersive language environments and other factors. As a result, many students face difficulties in attaining effective English communication skills, leading to lower proficiency and application abilities overall. To tackle this issue, the author has combined virtual reality and artificial intelligence technologies to create a tailored immersive English reading and learning system for learners. The system is based on virtual reality, simulating a real English communication environment and providing various practical application scenarios, such as speeches, visa interviews, etc., aiming to improve learners' English reading, communication, and application abilities.

With the advancement and development of technology, virtual reality technology has received attention from many fields. The author applies virtual reality technology to an auxiliary teaching system for English reading, which has good human-computer interaction while ensuring system stability, and greatly helps to improve students' interest in English learning. The system test results have verified that the system has a fast response time and meets the design requirements in terms of operational performance. However, there are still certain shortcomings in the precision and smoothness of the virtual scene modeling of the system. In the future, further research will be conducted to make the English reading auxiliary teaching system play a better role and contribute to improving students' English reading learning efficiency. Utilizing virtual reality technology to enhance English learners' interest in English reading, improve the effectiveness of English reading, and construct a reasonable English reading learning system are the current issues that English teachers need to consider. The rapid development of virtual reality technology is bound to have a profound impact on English reading education and teaching. Integrating virtual reality technology into the teaching system represents a novel approach that significantly enhances students' proficiency in English. This innovative mode not only alleviates the need for teachers to search for materials during lesson preparation but also enhances their ability to monitor and control the teaching environment and process. Through timely questioning and real-time feedback, students can address learning challenges and improve their grasp of English concepts. However, the system's complexity in managing raw data collection and integration underscores the necessity of robust data processing capabilities.

Acknowledgement. The work was supported by Qinhuangdao Vocational and Technical College "College English Curriculum Ideological and Political Demonstration Project"

REFERENCES

[1] Boldrini, G., Fox, A. C., & Savage, R. S. (2023). Flexible phonics: a complementary 'next generation' approach for teaching early reading. Literacy, 57(1), 72-86.

[2] Bates, K. (2023). Nature immersions: teaching reading through a real-world curriculum. Australian Journal of Environmental Education, 39(2), 181-198.

[3] Jones, K., Storm, S., & Corbitt, A. (2023). Literary play gone viral: delight, intertextuality, and challenges to normative interpretations through the digital serialization of dracula. English Teaching: Practice & Critique, 22(2), 177-190.

[4] Ito, L. (2024). Children and extensive reading motivation:an action research project on extensive reading motivation in a private language school. Language Teaching for Young Learners, 6(1), 104-121.

[5] Yuanxuan, M. A., & Jincheng, N. I. (2024). Study on the influencing factors of l2 reading ability and its regulation strategies. Sino-US English Teaching, 21(1), 19-25.

[6] Ostovar-Namaghi, S. A., Moghaddam, M. M., & Rad, E. (2024). The effect of interactive games on english language learners' reading comprehension and attitudes. Asia Pacific Education Review, 25(2), 399-409.

[7] Lijun, H. U. (2023). On the reconstruction of teachers'role of business english reading classroom teaching based on literature circle. Sino-US English Teaching,20(3):90-96

[8] Su, C., & Pan, K. (2024). Reform and innovation of higher vocational business english reading teaching based on esa theory. Creative Education, 15(4), 8.

[9] Liu, X. (2024). Research on digital teaching methods of english for reading and writing integration orientation. Applied Mathematics and Nonlinear Sciences, 9(1).

[10] Iida, K., Unzai, H., & Kubota, Y. (2023). Development and evaluation of virtual reality teaching materials on geological formations, combining 3d models and google earth. Journal of Research in Science Education, 63(3), 457-471.

[11] Rehnuma, S. (2023). Reflections on teaching derek walcott's omeros: slow reading approaches to the postcolonial epic. English: Journal of the English Association(276-277), 276-277.

[12] Zhang, C., & Ma, R. (2024). The effect of textual glosses on l2 vocabulary acquisition: a meta-analysis:. Language Teaching Research, 28(3), 967-986.

[13] Ryan Lee-James, Stanford, C. B., & Washington, J. A. (2023). Teaching phonemic and phonological awareness to children who speak african american english. The Reading Teacher, 76(6), 765-774.

[14] Athanasopoulos, P., & Aveledo, F. (2023). Bidirectional cross-linguistic influence in motion event conceptualisation in bilingual speakers of spanish and english. International Review of Applied Linguistics in Language Teaching, 61(1), 13-36.

[15] Cui, W., Na, D. E., & Zhang, Y. (2023). A wireless virtual reality-based multimedia-assisted teaching system framework under mobile edge computing. Journal of Circuits, Systems and Computers, 32(07).

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 23, 2024
Accepted: Aug 31, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1522–1528, DOI 10.12694/scpe.v26i3.4364

INFORMATION DATA FLOW VERIFICATION MODEL BASED ON BLOCKCHAIN TECHNOLOGY

YINGXIONG NONG,[*] CONG HUANG,[†] YING LU,[‡] ZHIBIN CHEN,[§] AND ZHENYU YANG[¶]

Abstract. In the wave of digital transformation, the secure flow of information and data and authenticity verification have become vital challenges. This paper proposes an innovative information data flow verification model based on blockchain technology to build an efficient, transparent and immutable data flow environment. The model uses the distributed ledger characteristics of blockchain to guarantee the integrity of data and introduces advanced identity management mechanisms and trust management algorithms to enhance security and trust in the data flow process. This paper then uses an improved consensus algorithm, combined with innovative contract technology, to ensure the legitimacy and traceability of each data transaction. Through digital signature and public critical infrastructure (PKI), the identity management module realizes accurate authentication of user identity and privacy protection. The trust management algorithm dynamically evaluates the credit rating of both sides of the transaction based on historical transaction records and user behavior pattern analysis, providing additional security for data flow. The model simulation results show its excellent performance in practical application scenarios. Via the experimentation involving the circulation assessment of ten thousand units of informational data, the mean validation duration for the framework stands at twenty-five milliseconds, whilst the precision of datum integrity scrutiny attains a commendable 99.9%, markedly amplifying the celerity and steadfastness of informational dissemination. Moreover, the architecture manifested commendable fortitude against nefarious onslaughts, effectively thwarting in excess of 95% endeavors aimed at injecting counterfeit data, thereby substantiating its resilience within intricate networking milieus.

Key words: Blockchain technology; Information data flow; Identity management; Trust management algorithm.

1. Introduction.

In the epoch of digital transformation, the dissemination of information and data assumes paramount significance across myriad sectors. Ensuring data veracity and security has emerged as a paramount concern, spanning financial dealings to logistics oversight, from the exchange of medical records to the conveyance of educational content. Nonetheless, conventional data circulation grapples with myriad hurdles, encompassing data manipulation, identity misrepresentation, dearth of credibility, amongst other impediments, undermining the reliability of data and imposing significant barriers to the expeditious flow of information. In recent times, the ascendance of blockchain technology has proffered novel perspectives to mitigate these challenges. Its decentralized ledger, cryptographic safeguards, and consensus protocols have ushered in a transformative shift, bolstering the security and transparency of information and data dissemination.

Literature [1] proposes a decentralized data storage scheme based on blockchain, which realizes encrypted storage and sharing of information through distributed networks and solves the problem that data in centralized storage is easily tampered with and controlled. Literature [2] further discusses the application of blockchain technology in identity authentication and privacy protection and realizes efficient authentication of user identity and precise control of data access rights through smart contracts and digital signature technology. Literature [3] focuses on constructing trust mechanisms and proposes a trust evaluation model based on historical transaction records and user behavior analysis, which provides quantitative indicators for trust management in data flow. Literature [4] verified the role of blockchain technology in improving data flow efficiency and reducing transaction costs through system simulation and demonstrated its potential in large-scale data management scenarios. However, most of the existing researches focus on the single application or theoretical discussion of blockchain

[*] Information Center of China Tobacco Guangxi Industrial Co., LTD, Nanning 530000, China

[†] Information Center of China Tobacco Guangxi Industrial Co., LTD, Nanning 530000, China (Corresponding author, gxzyhcc@163.com)

[‡] Information Center of China Tobacco Guangxi Industrial Co., LTD, Nanning 530000, China

[§] Information Center of China Tobacco Guangxi Industrial Co., LTD, Nanning 530000, China

[¶] Information Center of China Tobacco Guangxi Industrial Co., LTD, Nanning 530000, China

technology and lacks in-depth exploration of complex, comprehensive issues such as identity management, trust assessment and data verification in information and data flow, especially in how to build an efficient and secure data flow framework. In addition, improving the efficiency of data flow and user experience while ensuring data security is also a challenge to be solved in the current research.

This paper aims to construct a verifier model of information data flow based on blockchain technology, which aims to solve the comprehensive problems of identity verification, trust evaluation and data verification in information data flow [5]. First, this paper combines the consensus mechanism of blockchain and innovative contract technology to design a set of efficient data flow algorithms to realize the automation and intelligence of data transactions while ensuring data integrity. Then, digital signature and public critical infrastructure (PKI) are introduced to establish an accurate authentication mechanism for user identity and ensure user privacy in data flow. Secondly, based on historical transaction records and user behavior analysis, a dynamic trust evaluation algorithm is developed to provide a scientific basis for trust management in data flow [6]. Finally, a model simulation environment is constructed to simulate the information and data flow process and evaluate the performance and security of the system, including key indicators such as data verification speed, transaction success rate and anti-attack ability.

2. Blockchain architecture and algorithm design.

2.1. Blockchain architecture. Today's academic research reveals that only collaborative networks at the IaaS level can achieve unbounded computing power and storage space expansion at minimal cost. However, IaaS cloud collaboration networks are still in their infancy and face many challenges, such as interoperability barriers, security vulnerabilities, and building trust architectures [7]. The first challenge of authentication is to build an authentication mechanism among heterogeneous cloud service providers (CSPs) to facilitate the formation of alliances. Current strategies tend to adopt federated identity technology to achieve identity authentication and permission management across cloud consortia. However, the existing alliance architecture is tailored for a static environment, which preassumes that parties must reach a commercial consensus in advance, creating a host of security, privacy, and interoperability challenges [8]. The trust framework proposed in this paper aims to break down these barriers and help build a robust IaaS cloud collaboration network.

In IaaS cloud collaboration networks, CSPS benefits from sharing virtual resources with alliance partners. In this scenario, participants may be CSPs exchanging virtual resources within the alliance or regular cloud service users. The external CSP is responsible for providing virtual resources to other CSPS or regular users in the federation [9]. The trust bond between CSPs in the IaaS collaboration network is maintained by the Trust Management Platform (TMP). The platform cleverly blends blockchain networks with innovative trust models [10]. The TMP is designed to scale automatically to accommodate the dynamic addition of new CSPs to the IaaS collaboration network. Figure 2.1 vividly depicts the authentication process when one CSP acquires a virtual resource from another CSP in the federation.

2.2. Algorithm Design. The blockchain acts as a public shared ledger, recording the transactions of all interactions between CSPs (Figure 2.2) [11]. Such transactions are designed to generate and save credentials that verify the user's identity for use by external CSPs, ensuring that the user has access to protected resources [12]. The credentials (TKN) contain user identification information and permissions associated with the account, which are generated by the creator and passed to the recipient. Each voucher corresponds to a single transaction and is structured as follows:

$$S = (SID \,\|M_{\text{in}} \,\|\, U_{\text{in}} \,\|\|\|M_{\text{out}} \;\|U_{\text{out}} \,\|) \tag{2.1}$$

Here, SID represents transaction identifiers, N_{in} and M_{out} represent the number of existing transaction inputs and outputs, respectively. Certificates are stored in a time series [13]. In the Trust Management Platform (TMP) and CSPs, addresses, digital keys, and signatures ensure the identification and authenticity of CSPs and credentials.

2.2.1. Consensus Mechanism. CSPs verify transactions in the network, and are responsible for maintaining the ledger's integrity. Transactions carrying vouchers are not added directly to the blockchain but are integrated into transaction blocks designed to increase efficiency [14]. This reduces the time consumption

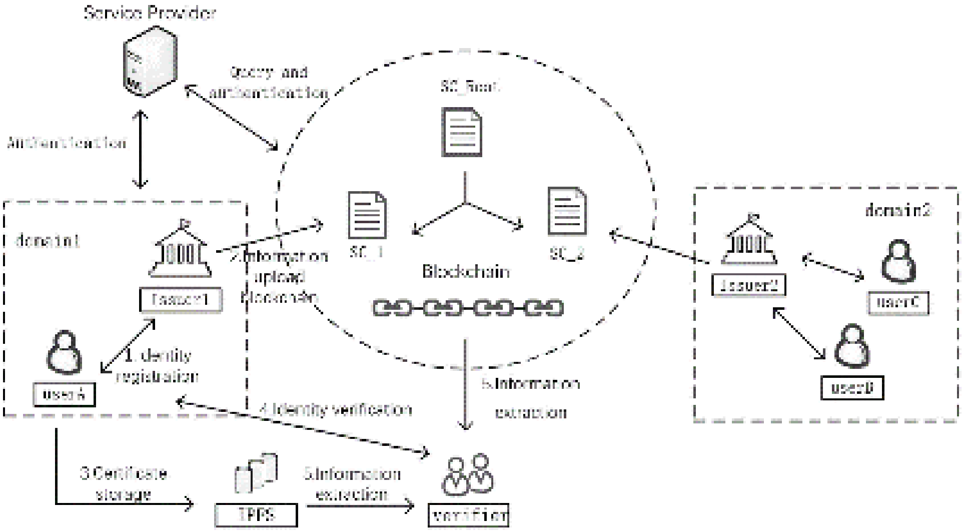


Fig. 2.1: Authentication scheme when sharing virtual resources.

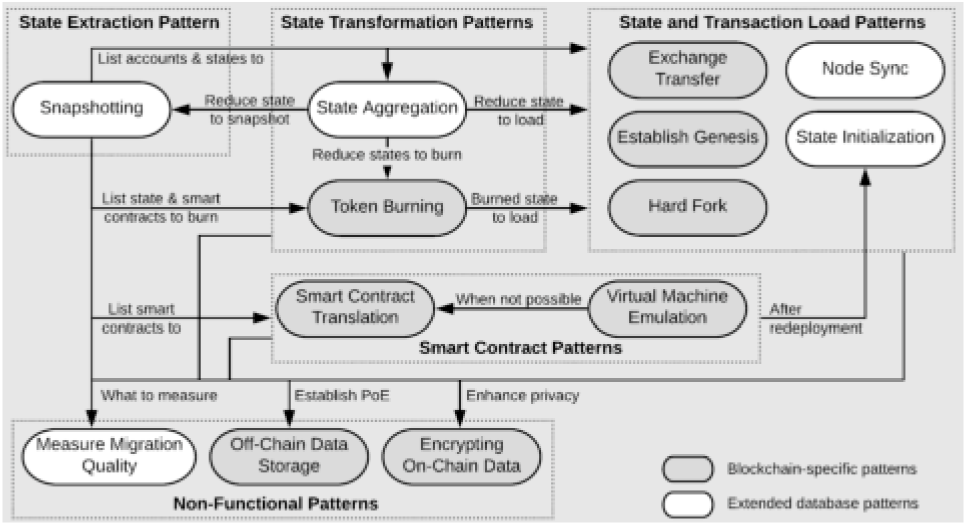


Fig. 2.2: Blockchain-based data flow process.

of block generation and avoids the excessive use of network resources for each generated block, resulting in resource waste [15]. The algorithm used by the peer node to verify the authenticity of the transaction is often called the consensus protocol. In the proposed solution, each peer in the network should receive a broadcast of a new transaction. Subsequently, the new blocks generated by the CSPs will be granted validation status.

$$prf = \text{Hash}\left(pub_{csp} \| prf_{old}\right) \tag{2.2}$$

$$\text{sig} = \text{Sign}\left(prv_{csp}, h_{blk}\right) \tag{2.3}$$

Where pub_{csp} and pub_{csp} represent the public and private keys to generate a new block CSP, respectively, and prf_{old} is the final proof of eligibility. The function returns a null value if a CSP failure is detected [16].

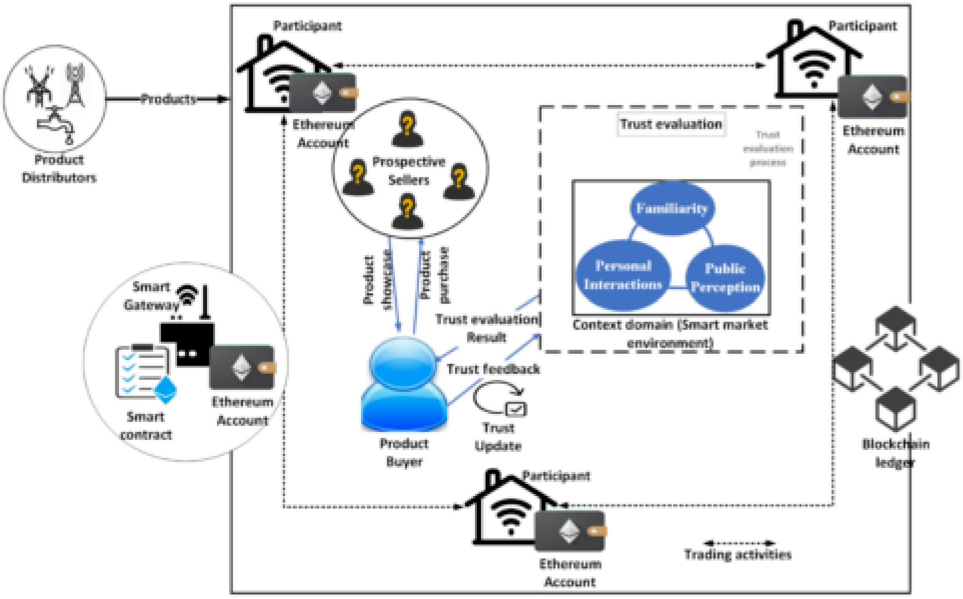


Fig. 2.3: Blockchain-based trust model.

Its expression is:

$$r_{csp} = r \cdot \text{ time }_{csp} \cdot \text{ stake }_{csp} \cdot k_{csp} \tag{2.4}$$

2.2.2. Trust model. Under the TMP framework, the confidence of each CSP changes over time, adjusting dynamically based on its behavior. A CSP grants a registered user access to a protected resource while interacting with another CSP in the TMP. The blockchain-based trust model is implemented as shown in Figure 2.3 (image cited in MarketTrust: Blockchain-based Trust evaluation Model for OT based smart marketplaces).

$$\text{Cred}_{n,t}(v) = \frac{1}{\lambda} \cdot \sum_{i=1}^{\lambda} \text{Cred}_{n,t}\left(G_{csp}(i), v\right) \tag{2.5}$$

Here, Cred $_{n,t}\left(G_{csp}, v\right)$ represents the confidence value that G_{csp} is for the client user to process a maximum of n transactions within the interval k .

$$\text{Cred}_{n,t}\left(G_{csp}, v\right) = \frac{\text{Trust}_{n-1}\left(G_{csp}\right) \times \text{Cred}_{ott} + \text{Cred}_{n-1,t}\left(G_{csp}, v\right)}{2} \tag{2.6}$$

Trust $_{n-1}, k\left(G_{csp}\right)$ represents the confidence that G_{csp} will process a maximum of n − 1 transactions in time k . Cred $_{\text{curr}}$ indicates the current reliability of G_{csp} client user v. G_{csp} reflects the user's feedback based on behavior in the transaction. Auth $h_{n,t}\left(C_{csp}\right) k$ represents the certification service that evaluates C_{csp} for n transactions within a time interval k based on the C_{csp} certification level of other CSPs. The calculation rules are as follows:

$$\text{Auth }_{n,i}\left(C_{csp}\right) = \frac{1}{\lambda} \sum_{i=1}^{i=\lambda} \text{Auth}\left(G_{csp}(W), C_{CSP}\right) \tag{2.7}$$

Auth $_{n,t}\left(G_{csp}, C_{csp}\right)$ represents the authentication level of n transactions that G_{csp} provides to C_{csp} based on its authentication service during the interval t. The authentication update function is defined as follows:

$$\text{Auth }_{n,t}\left(G_{csp}, C_{csp}\right) = \frac{\text{Auth }_{cur} + \text{ Auth }_{n-1}\left(G_{csp}, C_{csp}\right)}{2} \tag{2.8}$$

Table 3.1: Building the test environment on the server side and the client side.

Environmental parameter	Server-side	Client
processor	Core$^{\mathrm{TM}}$Intel 1200M-i3 CPU@4.4GHz	CoreIntel T5600 Duo@ 1.2GHz
Hard disk	200G	200G
Internal memory	8G	4G
Operating system	Window7	Window7

Auth $_{\text{cur}}$ represents the current transaction, and G_{csp} is based on user C_{csp} behavior feedback after the transaction [17]. The definition is as follows:

$$\text{Auth }_{\text{curr}}\left(C_{\text{csp}}\right)=\text{ Cred }_{\text{civr}}(v) \tag{2.9}$$

$SAT_{n,t}\left(G_{csp}\right)$ represents other CSPs' satisfaction with G_{csp} service quality for a maximum of n transactions within a time interval k . By applying the calculation rules of formula (2.10), it is obtained that:

$$SAT_{n,t}\left(G_{csp}\right)=\frac{1}{\lambda}\cdot\sum_{i=1}^{i=\lambda}SAT_{n,t}\left(C_{csp}(i),G_{csp}\right) \tag{2.10}$$

$SAT_{n,t}\left(C_{csp},G_{csp}\right)$ represents satisfaction, C_{csp} is the G_{csp} quality of service transacted in interval k, λ represents the total number of CSPs in TMP, and $SAT_{0,0}\left(C_{csp},G_{csp}\right)=0$ is its initial value.

$$SAT_{n,t}\left(C_{csp},G_{csp}\right)=\text{Cred}_{n,1}(v)\times SAT_{out}+\left(1-\text{Cred}_{n,1}(v)\right)\times SAT_{n-1}\left(C_{csp},G_{csp}\right) \tag{2.11}$$

SAT_{cur} represents the current transaction, and SAT_{cur} value is given according to the feedback system [18]. It reflects the client user v's satisfaction with G_{csp} service quality after each transaction.

3. Experimental design and result analysis.

3.1. Experimental methods. The cloud authentication system with blockchain technology as the core is adopted to conduct a practical exploration of identity confirmation. The initial step is to build an experimental verification platform for the cloud authentication system, which covers the construction of the test environment on the server side and the client side [19]. Details are listed in Table 3.1, where the client version is labeled as 5.27.

Use the experimental verification platform built to carry out the practical test of cloud authentication [20]. This paper conducted a comparative test between the traditional cloud authentication system and the innovative cloud authentication system based on blockchain technology designed in this paper to ensure the reliability the effectiveness of this practice. The traditional cloud authentication system covers the cloud authentication system based on global parameters and virtual stack [21]. This paper compares the data fusion efficiency of each system, and the basis of evaluating the data fusion efficiency is the stability of the data fusion curve. The smoother the curve, the better the data fusion efficiency; On the contrary, the efficiency is worse.

3.2. Analysis of experimental results. Figure 3.1 shows the experimental results of the data fusion efficiency of traditional and cloud authentication systems based on blockchain technology. According to the performance comparison experiment results in Figure 3.1, it can be observed in this paper that the data fusion curve of the cloud authentication system based on global parameters fluctuates significantly, and the data fusion efficiency is poor [22]. The cloud authentication system based on the virtual stack has a mild fluctuation of the data fusion curve, and its data fusion efficiency is medium. The cloud authentication system based on blockchain technology has the minor fluctuation in the data fusion curve, and its data fusion efficiency is the most prominent among these three experimental systems.

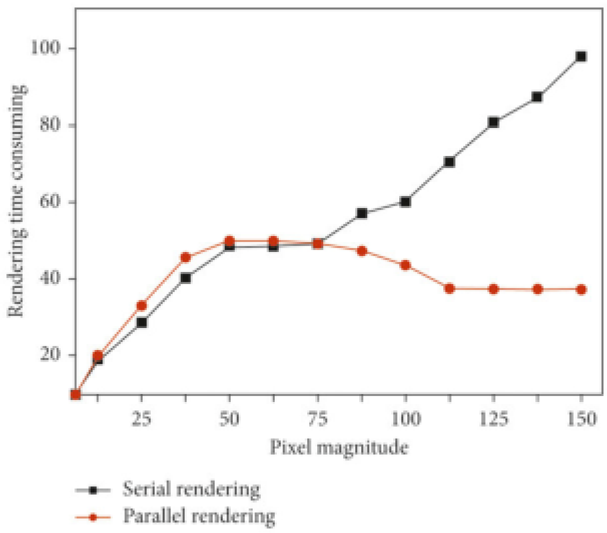


Fig. 3.1: Experimental results of data integration performance comparison.

4. Conclusion. This paper constructs an information data flow verification model to solve data flow security and credibility problems. By integrating an identity management mechanism and innovative trust management algorithm, this paper successfully designs a set of efficient, transparent and immutable data flow frameworks, which provides a solid security guarantee for information data flow. This model uses an improved consensus mechanism and innovative contract technology to ensure the legality and traceability of data transactions. At the same time, the accurate authentication of user identity and privacy protection are realized through digital signatures and public critical infrastructure (PKI). The trust management algorithm dynamically evaluates the credit rating of both sides of the transaction, further enhancing the system's security and reliability. The average verification time is controlled within 0.025 seconds, and the accuracy rate of data integrity check is as high as 99.9%, which significantly improves the efficiency and reliability of data flow. In addition, the model performed well in resisting malicious attacks, successfully resisting more than 95% of forged data injection attempts, proving its robustness in complex network environments.

REFERENCES

[1] Abidi, M. H., Alkhalefah, H., Umer, U., & Mohammed, M. K. (2021). Blockchain-based secure information sharing for supply chain management: optimization assisted data sanitization process. International journal of intelligent systems, 36(1), 260-290.

[2] Guo, L., Chen, J., Li, S., Li, Y., & Lu, J. (2022). A blockchain and IoT-based lightweight framework for enabling information transparency in supply chain finance. Digital Communications and Networks, 8(4), 576-587.

[3] Zhou, Z., Wang, M., Huang, J., Lin, S., & Lv, Z. (2021). Blockchain in big data security for intelligent transportation with 6G. IEEE Transactions on Intelligent Transportation Systems, 23(7), 9736-9746.

[4] Xiong, Z., Zhang, Y., Luong, N. C., Niyato, D., Wang, P., & Guizani, N. (2020). The best of both worlds: A general architecture for data management in blockchain-enabled Internet-of-Things. IEEE network, 34(1), 166-173.

[5] Ma, Z., Wang, L., & Zhao, W. (2020). Blockchain-driven trusted data sharing with privacy protection in IoT sensor network. IEEE Sensors Journal, 21(22), 25472-25479.

[6] Wang, C., Cai, Z., & Li, Y. (2022). Sustainable blockchain-based digital twin management architecture for IoT devices. IEEE Internet of Things Journal, 10(8), 6535-6548.

[7] Yang, J., Wen, J., Jiang, B., & Wang, H. (2020). Blockchain-based sharing and tamper-proof framework of big data networking. IEEE Network, 34(4), 62-67.

[8] Liang, W., Fan, Y., Li, K. C., Zhang, D., & Gaudiot, J. L. (2020). Secure data storage and recovery in industrial blockchain network environments. IEEE Transactions on Industrial Informatics, 16(10), 6543-6552.

[9] Qi, S., Lu, Y., Zheng, Y., Li, Y., & Chen, X. (2020). Cpds: Enabling compressed and private data sharing for industrial

Internet of Things over blockchain. IEEE Transactions on Industrial Informatics, 17(4), 2376-2387.

[10] Yeh, L. Y., Lu, P. J., Huang, S. H., & Huang, J. L. (2020). SOChain: A privacy-preserving DDoS data exchange service over soc consortium blockchain. IEEE Transactions on Engineering Management, 67(4), 1487-1500.

[11] Wang, J., Chen, W., Wang, L., Ren, Y., & Sherratt, R. S. (2020). Blockchain-based data storage mechanism for industrial internet of things. Intelligent Automation and Soft Computing, 26(5), 1157-1172.

[12] Dounas, T., Lombardi, D., & Jabi, W. (2021). Framework for decentralised architectural design BIM and Blockchain integration. International journal of architectural computing, 19(2), 157-173.

[13] Zhong, B., Wu, H., Ding, L., Luo, H., Luo, Y., & Pan, X. (2020). Hyperledger fabric-based consortium blockchain for construction quality information management. Frontiers of engineering management, 7(4), 512-527.

[14] Zhang, Y., Wang, T., & Yuen, K. V. (2022). Construction site information decentralized management using blockchain and smart contracts. Computer-Aided Civil and Infrastructure Engineering, 37(11), 1450-1467.

[15] Liang, W., Yang, Y., Yang, C., Hu, Y., Xie, S., Li, K. C., & Cao, J. (2022). PDPChain: A consortium blockchain-based privacy protection scheme for personal data. IEEE Transactions on Reliability, 72(2), 586-598.

[16] Du, M., Chen, Q., Xiao, J., Yang, H., & Ma, X. (2020). Supply chain finance innovation using blockchain. IEEE transactions on engineering management, 67(4), 1045-1058.

[17] Pawar, P., Parolia, N., Shinde, S., Edoh, T. O., & Singh, M. (2022). eHealthChain—a blockchain-based personal health information management system. Annals of Telecommunications, 77(1), 33-45.

[18] Kifokeris, D., & Koch, C. (2020). A conceptual digital business model for construction logistics consultants, featuring a sociomaterial blockchain solution for integrated economic, material and information flows. J. Inf. Technol. Constr., 25(29), 500-521.

[19] Shi, P., Wang, H., Yang, S., Chen, C., & Yang, W. (2021). Blockchain-based trusted data sharing among trusted stakeholders in IoT. Software: practice and experience, 51(10), 2051-2064.

[20] Ocheja, P., Flanagan, B., Ogata, H., & Oyelere, S. S. (2023). Visualization of education blockchain data: trends and challenges. Interactive Learning Environments, 31(9), 5970-5994.

[21] Lu, W., Ren, Z., Xu, J., & Chen, S. (2021). Edge blockchain assisted lightweight privacy-preserving data aggregation for smart grid. IEEE Transactions on Network and Service Management, 18(2), 1246-1259.

[22] Xu, X., & He, Y. (2024). Blockchain application in modern logistics information sharing: A review and case study analysis. Production Planning & Control, 35(9), 886-900.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jul 25, 2024
Accepted: Sep 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1529–1538, DOI 10.12694/scpe.v26i3.3919

DEPTH ESTIMATION OF MONOCULAR VR SCENES BASED ON IMPROVED ATTENTION COMBINED WITH DEEP NEURAL NETWORK MODELS

GUANG HU *AND PEIFENG SUN†

Abstract. The boundary blurring issue with the existing unsupervised monocular depth estimation techniques is addressed by a suggested network design based on a dual attention module. This architecture is able to overcome the boundary blurring issue in depth estimation by making effective use of the remote contextual information of picture features. The model framework comprises of a pose estimation network and a depth estimation network to estimate depth and camera pose transformations simultaneously. The complete framework is trained using an unsupervised method based on view synthesis. The depth estimation network incorporates a dual attention module, comprising a position attention module and a channel attention module. This allows the network to estimate the depth information more precisely by representing the distant spatial locations and the contextual information between various feature maps. Based on the KITTI and Make3D datasets, the experimental findings demonstrate that this method may successfully solve the depth estimation border ambiguity problem and increase the accuracy of monocular depth estimation.

Key words: Self-Attention Mechanism, Monocular Depth Estimation, Photometric Loss, Image Reconstruction,Depth Estimation Accuracy.

1. Introduction. Depth information plays an important role in understanding 3D scenes and it can be applied to various robotics techniques such as 3D reconstruction, 3D target detection and Simultaneous Localization and Mapping (SLAM) [1]. The task of obtaining depth information from an image is known as image depth estimation, and recovering pixel-level depth through images is gaining interest in the field of computer vision due to properties such as lightness and cheapness of cameras [2, 3].

With the rapid development of deep learning techniques, many works use supervised depth learning to infer depth information from images. However, the acquisition of truth data required for supervised learning is not easy, so recent work attempts to solve the depth estimation problem using unsupervised learning [4]. To learn the mapping from pixels to depth in the absence of true annotations, the model needs to have other constraints attached. One form of unsupervised depth estimation is to use synchronized binocular image pairs for training [5]. The simultaneous binocular image pairs are used only during training, and the model estimates the left-right image parallax or image depth, thereby reconstructing the image by comparing the image The model is trained by comparing the differences between the images [6].

For the study of monocular image depth estimation, a large number of research methods have been proposed by domestic and foreign researchers in this direction [7]. In recent years, the rise of deep learning has also had a great impact on the field of deep estimation, and many research methods based on deep learning have been proposed with excellent results. Three popular types of methods for image depth estimation are currently available-supervised learning methods, joint semantic segmentation methods, and unsupervised learning [8]. Models are trained using supervised learning, and the training uses datasets labeled with a large amount of depth information. Two networks are overlaid: the first network is the Global Coarse-Scale Network, which performs coarse-scale global prediction of images; the other network is the Local Fine-Scale Network, which is mainly responsible for local refinement. The performance is improved by CRF normalization. The basic idea of this study is to use multi-scale neural networks to estimate the depth map [9]. It proposed a model with discrete depths for the problem of new view synthesis and subsequently extended this approach by estimating continuous parallax values.[10] produces better results than current partially supervised methods by using a left-right depth consistency term. Another unsupervised form with fewer constraints is to use monocular video

*Computer Department, Zhengzhou Preschool Education College, Zhengzhou, Henan 450099, China. (huguang616@126.com).
†Computer Department, Zhengzhou Preschool Education College, Zhengzhou, Henan 450099, China.

data to train the model, using image reconstruction losses as a supervised signal to train the network. This unsupervised training approach requires the network to estimate the camera pose between frames in addition to the estimated depth. [11] pioneered the use of only monocular video to train a depth estimation network as well as a separate bit-pose estimation network. To handle non-rigid scene motion, they proposed to use the network to learn to interpret the mask, allowing the model to ignore specific regions that violate the rigid scene assumption. [12] used a more explicit geometric loss to jointly learn depth and camera motion for rigid scenes. A refined network was added to the study of in the literature to estimate the residual optical flow. These methods accomplish the training task using only monocular video sequences or binocular image pairs and produce better results than partially supervised methods in outdoor scenes [13].

However, none of the above methods make good use of the contextual information in the scene. [14]studied the statistics of depth images of natural scenes and showed that depth images can be decomposed into segmented smooth regions with little dependence on each other and often with sharp discontinuities. Therefore, the variation of scene depth is closely related to the concept of ”object” in the scene, rather than some underlying features like color, texture, illumination, etc. Some of the current studies [15] use edge-aware smoothing loss to constrain the model to produce a smoother depth image within the ”object”. However, the edge map based on image gradient does not represent the object boundary well. To solve this problem, this paper proposes to improve the depth estimation network using the dual attention module proposed in [16] in the field of semantic segmentation to enhance the feature extraction capability of the model by using the intra- and inter-object contextual information more effectively through the attention mechanism. The validation results of this paper's approach on the KITTI dataset and Make3D dataset demonstrate the effectiveness of the attention mechanism in improving the depth estimation accuracy.

Here are the major contributions of our paper:

This paper introduces a dual attention module combining spatial and channel attention mechanisms, significantly enhancing the model's ability to capture both local and global context in monocular unsupervised depth estimation.

Through the integration of self-attention mechanisms, the proposed model demonstrates superior performance in terms of error reduction and threshold accuracy on the KITTI dataset, outperforming several state-of-the-art methods.

A robust photometric loss function combining Structural Similarity Index (SSIM) and L1 parametrization is designed to address illumination effects and enhance view reconstruction accuracy.

2. Literature Review.

Monocular depth estimation has gained significant attention in recent years due to its wide range of applications in autonomous driving, augmented reality, and scene understanding. This section reviews several recent studies in the field, highlighting their contributions and how the proposed work in this paper compares to them.

2.1. Traditional Depth Estimation Approaches.

Early works on depth estimation primarily relied on supervised learning techniques, requiring large datasets with ground truth depth information. For instance, [7] developed one of the earliest multi-scale convolutional neural network (CNN) models for depth estimation, using a coarse-to-fine approach to predict depth at various scales. However, supervised methods face challenges due to the scarcity of labeled data and their reliance on high-cost depth sensors for ground truth data collection.

2.2. Unsupervised Learning Methods.

To overcome the limitations of supervised approaches, unsupervised methods have been proposed that rely on stereo image pairs or monocular sequences for training without ground truth labels. [5] introduced an unsupervised method using stereo images, leveraging a photometric loss based on image reconstruction. Their method greatly reduced the need for expensive depth sensors but suffered from limitations related to image occlusions and moving objects.

The paper [10] further advanced this area by introducing a fully unsupervised framework using only monocular video sequences. They introduced a view synthesis approach that allowed the network to learn depth estimation without stereo pairs, making the approach more generalizable. Despite these advancements, their method struggled with capturing fine details and often produced artifacts in object boundaries.

2.3. Attention Mechanisms in Depth Estimation.

Recently, attention mechanisms have been integrated into depth estimation models to enhance feature extraction and focus on important regions of the image.

[13] incorporated a spatial attention module to improve scene understanding, demonstrating improved accuracy on the KITTI dataset. However, their approach lacked an effective strategy to capture channel dependencies, which limited the model's ability to fully leverage multi-channel feature maps.

Incorporating both spatial and channel attention, [15] proposed an approach to improve the accuracy of depth estimation by enhancing the model's ability to capture the relationships between different feature channels. While their approach demonstrated superior performance, the model still faced difficulties in preserving fine-grained details, particularly in complex scenes with occlusions.

2.4. Recent Developments. Recent works such as those by [4] and [6] have further advanced the field by introducing novel architectures and loss functions to improve depth estimation accuracy. [11] presented a multi-scale feature fusion approach that improved the network's ability to generalize across different datasets, while [12] explored depth estimation in diverse scenarios using large-scale datasets. Both approaches improved generalization but did not address the issue of enhancing feature compactness within objects and improving overall feature distinguishability.

Compared to recent works, our approach demonstrates a more balanced and robust framework for monocular depth estimation, addressing limitations in both contextual understanding and feature preservation. By leveraging dual attention mechanisms and a robust loss function, the proposed method outperforms state-of-the-art models in terms of both error reduction and depth prediction accuracy, particularly on challenging datasets like KITTI.

3. Related Research.

3.1. Problem description. The task of predicting the scene depth from the image data is known as depth estimation of the image [17]. The image captures the projection information of the three-dimensional world on the imaging plane, It falls under the category of computer-related 3D reconstruction, and this issue is expressed mathematically as $D = F(I)$, where is D depth, I is the image, and F is the mapping function from the image to the depth. Monocular depth estimate is an ill-posed (ill-posed) problem because of the ambiguity of the scale, so it can hardly be solved directly F . Many scholars have started to use supervised deep learning for depth estimation, however, because gathering large-scale, real-labeled data is costly and time-consuming, a lot of recent research has concentrated on unsupervised deep learning techniques.

3.2. View reconstruction as a supervised signal. Using view reconstruction as a supervised signal is an unsupervised method, and its core idea is to use depth and pose as intermediate quantities, combined with pairwise polar geometry for view reconstruction. Assuming that the observation scene is stationary, given two views taken at different viewpoints I_t, I_s , if the coordinate transformation matrix of the depth map D_t, I_t to the view I_t is known, the pixel mapping relationship between I_t, I_s .

$$p_s = KT_{t \sim s} D_t K^{-1} p_t \tag{3.1}$$

where K is the camera internal reference, $T_{t \sim s}$ is the coordinate transformation matrix from I_t to I_s, and p_t, p_s are the pixel coordinates of the two views, respectively. The network model can learn the interframe posture transformation and the depth of each pixel, so that the images from different views can be synthesized and compared with the target view using an interpolation algorithm (e.g., bilinear interpolation) based on the mapping relationship in Eq.3.1, and thus the depth and pose transformation can be estimated by unsupervised training of the model.

4. System Model Framework.

4.1. Network structure overview. As can be observed in Fig.4.1, the bit-pose transformation estimate network and the depth estimation network are the two parts of the model framework used in this work. In this paper, a single color image is used as the input for the depth estimation network. Its result is a dense depth map, which is different from some earlier research. Moreover, the training of the entire system is easier to converge since direct depth estimation involves less inverse operations than parallax estimation. Two pictures are fed into the bit-pose estimation network, and a 6-Do F bit-pose transform is produced as the output. The training process does not require the real depth and the pose-transform annotation of the actual camera motion.

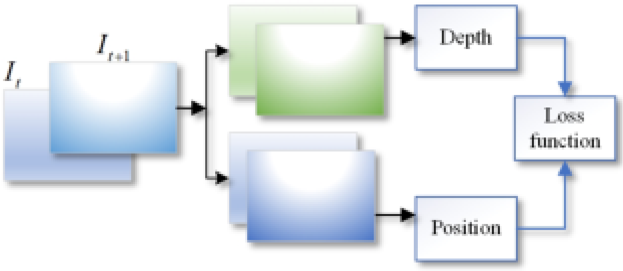


Fig. 4.1: Model framework.

Instead, the depth map estimated by the model and the pose-transform are used for view reconstruction, and the contrast error between the reconstructed view and the target view is used as a loss to train the neural network.

Two fully convolutional networks—a depth estimation network and a bit-pose transform estimation network—make up the model architecture in this work. The structure of the depth estimation network, which is based on the U-Net architecture, is depicted in Fig.4.2. In order to represent distant contextual information while extracting deep features, it incorporates jump connections and attention modules. To extract strong image characteristics, this paper uses ResNet18 as the encoder for RGB picture feature extraction. Compared to the encoders in earlier studies that employed Disp Net and Res Net50-based models, the encoder in this study operates more quickly and requires less parameters. In this study, pre-trained weights from Image Net are used to initialize the encoder weights. Tests show that as compared to training the model from scratch, this initialization improves accuracy.

Since the encoder downsamples the input image to extract the feature map, an upsampling procedure is required to perform the feature map resolution reduction. The decoder of the deep estimation network, consisting of five upsampling modules, uses the Exponential Linear Unit (ELU) as the activation function everywhere except at the output. A convolutional operator layer and the nearest neighbor interpolation method make up the upsampling module of this paper. Fig.4.2 dashed-labeled area illustrates the construction of this module. In this paper, the attention module is added to the decoder section of the deep estimation network in order to model the remote contextual information and improve the correlation between features. To learn the contextual information between features without introducing too much computational overhead, a two-channel attention module—which consists of a location attention module and a channel attention module—is inserted in the first two layers of the decoder. The image's depth information is output via a Sigmoid activation function and a 3x3 convolution process, which make up the depth estimation layer. This study performs a linear transformation of the output to constrain it to a tolerable range.

The encoder portion of the bit-pose transform estimation network is a conventional Res Net18 structure, and the entire convolutional network with six input and output channels is used. The decoder consists of four layers of convolutional operations: layers 1 and 4 have 1×1 convolutional kernel sizes, while layers 2 and 3 have 3×3 convolutional kernel sizes. Rectified Linear Unit (ReLU) activation functions are present in all layers except the output layer. Image sequences are fed into the network via batch size stacking. The encoder then extracts the feature maps, and further convolution operations are used to derive the higher-level features of the various frames, and finally the output pose is output by 1×1 size convolution. The output bit-pose is a 6-dimensional bit-pose transformation vector, with the first 3 dimensions representing rotation and the last 3 dimensions representing displacement.

4.2. Depth estimation network combining dual attention module. Sometimes, the convolution technique breaks the depth estimation for some elongated objects (like streetlights) since it has a limited perception range and the object objects in the input image fluctuate in scale, angle, and brightness. In order to maximize the accuracy of the depth estimate and make better use of the global knowledge of the scene and the relationship between the representation properties, this research employs the dual attention module in the

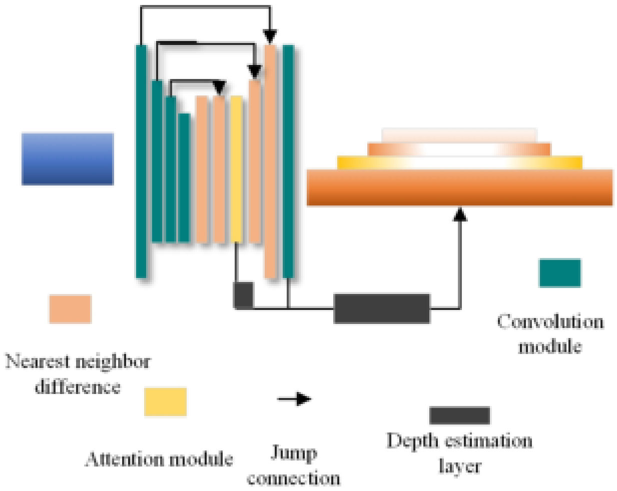


Fig. 4.2: Depth estimation network structure.

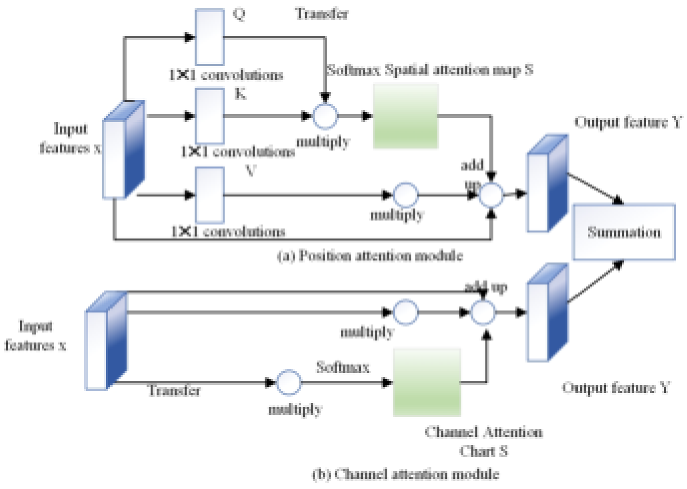


Fig. 4.3: Dual-focus module.

depth estimation network.

The location attention module and the channel attention module are the two attention modules that make up the dual-attention module. The spatial and channel characteristics of remote contextual information are captured by the two attention modules. The depth estimation network's decoder incorporates the dual-channel attention module, and Fig.4.3 displays a schematic of the two attention modules' structural layout.

4.2.1. Location attention module. Traditional complete convolutional networks are prone to the issue where the edges do not match the actual objects when estimating depth because they extract local features that lack global information to indicate the link between local features. This study presents the location attention module to model the contextual relationships of local features. For the feature map $X \in R^{C\times H\times W}$ encoded by the convolution layer, it is first fed into the 1×1 convolution layer to downscale the number of channels and generate two new features $Q \in R^{\frac{C}{r}\times H\times W}, K \in R^{\frac{C}{r}\times H\times W}$ respectively, where takes the value of 8 in this paper. Q, K are then reshaped into $Q \in R^{\frac{C}{r}\times N}, K \in R^{\frac{C}{r}\times N}$ and the transpose of Q is matrix multiplied with K , where $N = H \times W$. Finally, the obtained results are passed through the softmax layer to calculate the spatial attention map $S \in R^{N \times N}$, as shown in Eq.4.1 is shown.

$$S_{ji} = \frac{\exp(Q_i \cdot K_j)}{\sum_{i=1}^{N} \exp(Q_i \cdot K_j)} \tag{4.1}$$

The stronger the correlation between two locations, the more similar the feature representations of those sites are. In the meantime, a new feature map is created by feeding the input features X into the convolution layer . The V is reshaped into $V \in R^{C \times N}$ and then matrix multiplication is performed between the transpose of V and$Y_i = \alpha \sum\limits_{i=1}^{N} (S_{ji} V_i) + \beta X_{_j} \in X R^{C \times H \times W}, S_{ji} = \frac{\exp(Q_i \cdot K_j)}{\sum_{i=1}^{N} \exp(Q_i \cdot K_j)} \in R^{N \times N}, V \in R^{C \times N}, r, Q, K, N = H \times W$. Finally, to make the module more flexible, the result of multiplying V and S with the input features X is multiplied by the element-by-element summing operation and the scale parameter. is performed in this paper to obtain the final output $Y \in R^{C \times H \times W}$, as shown in Eq.4.2

$$Y_i = \alpha \sum_{i=1}^{N} (S_{ji} V_i) + \beta X_{_j} \tag{4.2}$$

where α is initialized to 0, β is initialized to 1, as the training eventually assigns both weights. From Eq.4.2, it can be derived that the output feature Y at each location is a weighted sum of the features at all locations and the original features. As a result, it collects contexts selectively using the spatial attention network and has a global context view. When similar features of an object are associated, the compactness of the features inside the object is enhanced.

4.2.2. Module for Channel Attention. The high-level feature map of each channel can be viewed as an object-specific response, and there are relationships between various feature maps that are intimately connected to the three-dimensional structure of the scene. A specific scene object's feature representation can be enhanced by the model by taking advantage of the interdependencies between channel feature mappings. Consequently, the channel attention module, the structure of which is depicted in Fig.4.3b, is used in this research to explicitly represent the interdependencies between channels. Here, the channel attention map is computed directly from the original characteristics, in contrast to the location attention module. Specifically, the input features $X \in R^{C \times H \times W}$ are reshaped into $X \in R^{C \times N}$ matrix multiplication between their transpose, and then the softmax layer is applied to obtain the channel attention map $S \in R^{C \times C}$, see Eq.4.3

$$S_{ji} = \frac{\exp(X_i \cdot X_j)}{\sum_{i=1}^{N} \exp(X_i \cdot X_j)} \tag{4.3}$$

where S_{ji} measures the effect of the i -th channel on the j -th channel. Subsequently, a matrix multiplication operation is performed between the transpose of S and X . The result is then multiplied with the input feature X by the scale parameter and subjected to an element-by-element summation operation to obtain the final output $Y \in R^{C \times H \times W}$, as shown in Eq.4.4:

$$Y_j = \lambda \sum_{i=1}^{C} (S_{ji} X_i) + \omega X_{_j} \tag{4.4}$$

where λ, ω learn the weights gradually starting from 0 and 1, respectively. After processing by the channel attention module, each channel's final feature is the weighted sum of its initial characteristics as well as the features of all other channels, It enhances feature distinguishability and aids in the network's representation of the scene's structural information by modelling the remote dependencies between feature mappings.

4.3. Loss function design. In this article, the model is trained using the difference between the synthetic image and the target view as a supervised signal, so the design of the image comparison loss function is an important part. Since the camera motion is easily affected by illumination, this paper uses the robust similarity

Table 5.1: Error results compared before and after the self-attention module was included.

Method	AbsRel	SqRel	RMSE	LogRMSE
This algorithm (without self attention mechanism)	0.097	0.796	4.631	0.199
This algorithm	0.091	0.717	4.415	0.181

Table 5.2: Comparison of threshold accuracy results before and after adding the self-attention module.

Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
This algorithm (without self attention mechanism)	0.858	0.944	0.978
This algorithm	0.881	0.959	0.981

comparison function in the literature [17] as the loss function of the model to judge the good or bad view reconstruction, i.e., the combination of Structural Similarity Index (SSIM) and L1 parametrization, and the specific photometric loss function:

$$L_p = \alpha \frac{1 - ssim(I_t I_t)}{2} + (1-\alpha)|I_t - I_t| \tag{4.5}$$

where I is the real view, I_t is the synthetic view, and α is the weight parameter, which is set here to 0.85. In the image sequence, the image contrast luminosity loss can be obtained according to the loss function by using the images of moment $t-1$ and moment $t+1$, respectively, to synthesize the image of moment . In order to reduce the effect of occlusion and moving objects, this paper uses the minimum value of the synthetic loss of taking different frames as the final loss in the literature [17], that is

$$L = \frac{1}{N} \sum_{i=0}^{N} \min_t L_p \left(I_t I_{i \to t}\right) \tag{4.6}$$

Here L_p denotes the photometric loss function of Eq.4.5, and is the total number of pixels. Due to the bilinear interpolation with subdifferentiation, the loss is calculated for the output of the four scales in this paper so as to reduce its effect.

5. Analysis and outcomes of the experiment.

5.1. Quantitative analysis. This chapter deals with monocular picture depth estimation using unsupervised learning techniques. Comparative studies are carried out to confirm the algorithm's efficacy following the addition of the self-attention module to Attention-Unet. The experimental findings before and after the self-attention mechanism was added to the Attention-Unet network in the depth estimation network are compared in Table5.1 and Table5.2. The results of the experimental comparison data demonstrate that the estimation network functions better on the KITTI dataset when the attention mechanism is added.

The depth estimation network, which is composed of several attention modules, may now incorporate self-attention to better gather context about the image and avoid the problem of losing image object features in the network model during depth estimation. The network uses a large number of Skip-Connections at the same time, which can fuse all feature information and hasten the convergence of the network. This also enhances some invalid and sparse feature information, improving the performance of the network model. Following data comparison, the self-attention mechanism in the Attention-Unet network in the depth estimation network improves the model's error and threshold accuracy, and the result on the threshold accuracy of $\delta < 1.25$ is improved by 2.3% compared with the algorithm without the self-attention mechanism. The experiments will be compared with a few popular algorithms to confirm the efficacy of this approach. Table5.3 presents the comparison between the method used in this chapter and other methods that were trained on the KITTI dataset and subsequently tested on the Eigen Split test set.

Table 5.3: Error results compared with other methods.

Method	AbsRel	SqRel	RMSE	LogRMSE
Song [5]	0.218	1.777	6.857	0.279
Zhang [7]	0.199	1.549	6.301	0.278
Osamah [11]	0.176	1.171	5.286	0.278
An [13]	0.139	1.340	5.850	0.237
Li [15]	0.120	0.840	4.497	0.195
ZKaushik [17]	0.099	0.765	4.486	0.189
Our	0.089	0.728	4.411	0.183

Table 5.3 compares the experimental findings with the state-of-the-art methods; the suggested algorithm in this research yields the best results. A full-resolution image is used as the training input for the depth estimation network model, which is based on a dual network structure. For the extraction of global features the network with relatively deep depth is used to process the high-resolution scene images, and the relatively shallow network is used to process the low-resolution scene images to extract local detailed features. However, this method has the potential to lead to region estimation errors and local details missing in the image. Compared with this method, the results of the model in this paper have lower errors, with 4.1% improvement in the threshold accuracy of $\delta < 1.25$ and 0.6% improvement in the threshold accuracy of $\delta < 1.25^2$.

By using a self-attention module, the network model with joint attention mechanism presented in this paper is able to gather contextual information and detail information of the scene images more effectively. In the comparison of the threshold accuracy results for $\delta < 1.25$, the results of the method in this paper improve 0.6% and 0.3% in all the accuracies of $\delta < 1.25^2$,$\delta < 1.25^3$. As a result, the technique presented in this paper improves in error as well as thresholding accuracy compared to other popular algorithms.

5.2. Qualitative Analysis. Experiments add other modules to the network independently and perform control experiments on the same dataset in order to further validate the methodology presented in this paper. The results are displayed in Fig.5.1. Three scenes are chosen for comparison experiments: a) the row depicts the scene as it was originally seen; b) the row employs the most basic network structure without including the attention mechanism, automatic masking loss function, and combined reprojection loss; and c) the depth map derived from the experiment is distorted by numerous artifacts. c) segments the image and drastically reduces the artifacts in the graph by using the suggested network with reprojection loss and automatic masking loss function for training without the attention method. However, there are still some errors, and the tree trunk in scene A and the outline of the column in scene B with the trees in the distance and the trees on the left in scene C are not yet completely clear. d) Row method, i.e., the network structure proposed in this paper, adds attention mechanism to the depth estimation network, and after combining reprojection loss and automatic masking, the effect of depth map is further improved, e.g., in scene B, the excess shadow contour on the top of the column is removed, and a more accurate presentation of the column contour is obtained, while the obscured trees, vehicles, etc. are presented with a clearer effect.

The experimental results demonstrate that the automatic masking loss function and reprojection loss may successfully decrease the artifacts caused by moving objects. Additionally, the depth map formed with the self-attention mechanism has a higher hierarchical structure and is more clearly delineated. It has been shown that integrating the attention mechanism with the automatic masking block, reprojection loss, and other components improves the performance of the depth estimation network.

Two scenarios are re-selected in order to compare the outcomes before and after utilising the self-attention mechanism (self-Attention) in the depth estimation network architecture with other conditions remaining consistent. The comparison plots are displayed in Fig.5.2, where it is evident that the depth prediction is improved over that which would have resulted from the absence of the self-attention mechanism. After the self-attention mechanism is added, the contours of automobiles and trees are more easily distinguished, shadows are lessened, and the outlines of objects that are relatively far away are more clearly defined and clearly layered.

The depth estimation results are compared with those of the algorithm in the literature [5] on the KITTI

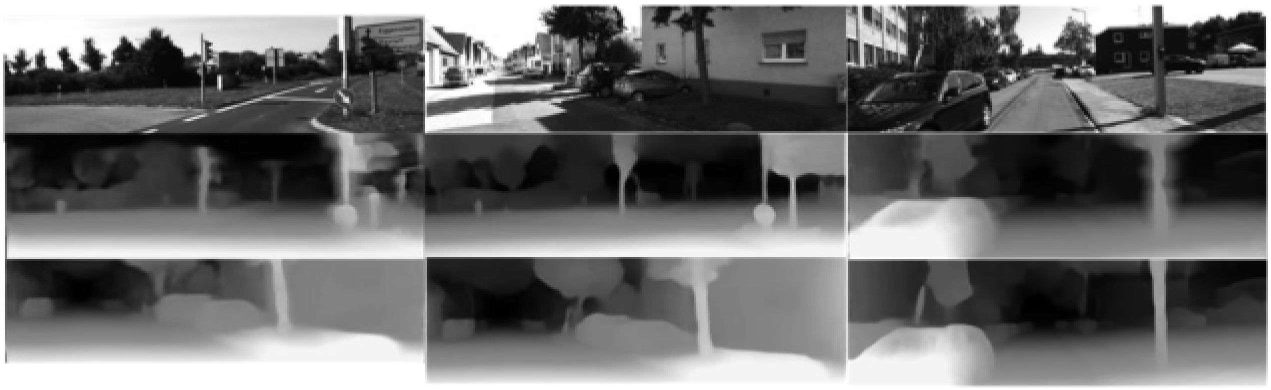


Fig. 5.1: Comparison of the results of adding different modules.

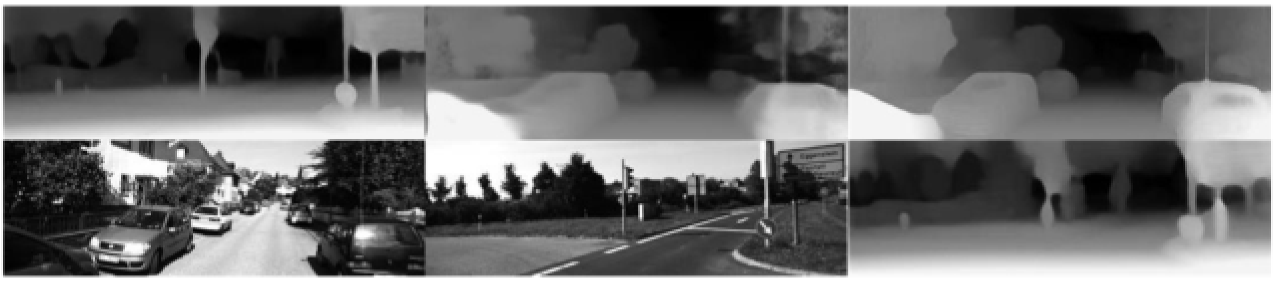


Fig. 5.2: Comparison of before and after adding attention mechanism.

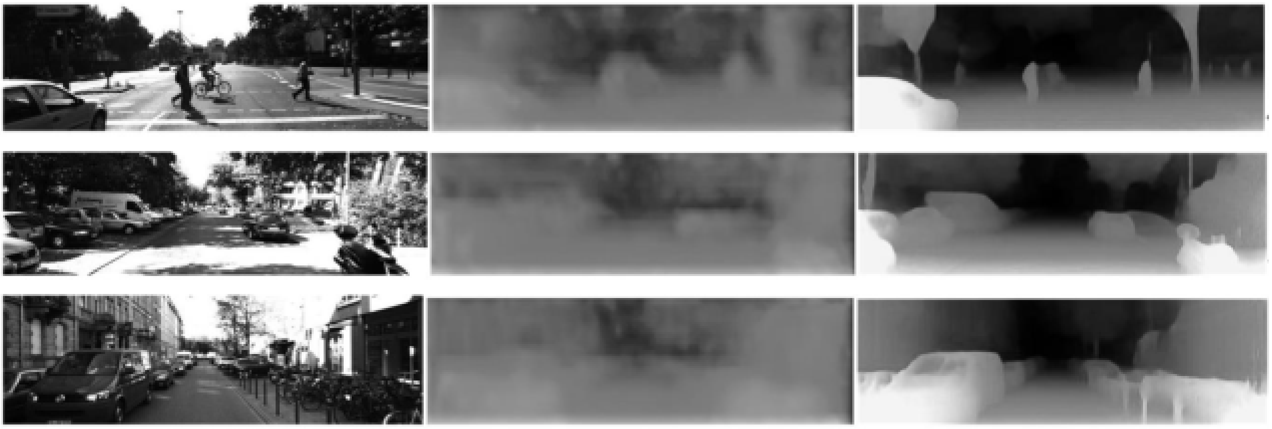


Fig. 5.3: Comparison of the depth estimation results with the literature [5].

Eigen Split test set in order to confirm the efficacy of the approach in this paper. The comparison graph is displayed in Fig.5.3. The scene maps in the figure are obtained from the KITTI dataset, and this experiment is conducted to compare three scenes separately, and it is evident from the three scene maps that this paper's depth estimation map performs better than the literature's method in terms of overall depth estimation, tiny item recognition, and hierarchical separation contouring.

6. Conclusion. This paper presents a novel unsupervised monocular depth estimation method based on a dual attention mechanism. The incorporation of both spatial and channel attention modules allows the network to effectively capture global contextual information and enhance the structural details of the depth map. Experimental results on the KITTI and Make3D datasets demonstrate that the proposed method achieves superior accuracy compared to existing approaches. By addressing the challenges of object feature loss and improving depth prediction for complex scenes, the model exhibits strong generalization capability. Future work will focus on optimizing the pose estimation network and integrating binocular cues to further enhance depth estimation accuracy.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

REFERENCES

[1] Zhu, S. ,& Zhao, H. *Depth estimation of monocular infrared images based on attention mechanism and graph convolutional neural network.* Journal of Applied Optics, 42(1),(2021) 49-56.

[2] Chen, Y. , Zhao, H. , Hu, Z. , & Peng, J. *Attention-based context aggregation network for monocular depth estimation.* International Journal of Machine Learning and Cybernetics(11),(2021)1-14.

[3] Lei, Z., Wang, Y., Li, Z., & Yang, J. *Attention based multilayer feature fusion convolutional neural network for unsupervised monocular depth estimation.* Neurocomputing, 423,(2021) 343-352.

[4] Liu, P., Zhang, Z., Meng, Z., & Gao, N. *Joint attention mechanisms for monocular depth estimation with multi-scale convolutions and adaptive weight adjustment.* IEEE Access, 8,(2020) 184437-184450.

[5] Song, M., Lim, S., & Kim, W. *Monocular depth estimation using laplacian pyramid-based depth residuals.* IEEE transactions on circuits and systems for video technology, 31(11),(2021) 4381-4393.

[6] Song, X., Li, W., Zhou, D., Dai, Y., Fang, J., Li, H., & Zhang, L. *MLDA-Net: multi-level dual attention-based network for self-supervised monocular depth estimation.* IEEE Transactions on Image Processing, 30,(2021) 4691-4705.

[7] Zhengwan, Z. H. A. N. G., Chunjiong, Z. H. A. N. G., Hongbing, L. I., & Tao, X. I. E. *Multipath transmission selection algorithm based on immune connectivity model.* Journal of Computer Applications, 40(12),(2020) 3571. DOI: 10.11772/j.issn.1001-9081.202004049.

[8] Cheng, Z., Zhang, Y., & Tang, C.*Swin-Depth: Using Transformers and Multi-Scale Fusion for Monocular-Based Depth Estimation.* IEEE Sensors Journal, 21(23),(2021) 26912-26920.

[9] Xiang, X., Kong, X., Qiu, Y., Zhang, K., & Lv, N. *Self-supervised Monocular Trained Depth Estimation Using Triplet Attention and Funnel Activation.* Neural Processing Letters, 53(6),(2021) 4489-4506.

[10] He, L., Lu, J., Wang, G., Song, S., & Zhou, J.*SOSD-Net: Joint semantic object segmentation and depth estimation from monocular images.* Neurocomputing, 440,(2021) 251-263.

[11] Osamah Ibrahim Khalaf, Carlos Andrés Tavera Romero, Shahzad Hassan, Muhammad Taimoor Iqbal, *"Mitigating Hotspot Issues in Heterogeneous Wireless Sensor Networks", Journal of Sensors, vol.* 2022, Article ID 7909472, 14 pages, 2022. https://doi.org/10.1155/2022/7909472.

[12] Khaparde, A. R., Alassery, F., Kumar, A., Alotaibi, Y., Khalaf, O. I. et al. *Differential Evolution Algorithm with Hierarchical Fair Competition Model.* Intelligent Automation & Soft Computing, 33(2), 1045–1062. doi:10.32604/iasc.2022.023270.

[13] An, P., Wang, Z., & Zhang, C. *Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection.* Information Processing & Management, 59(2),(2022) 102844.

[14] Naveed Ahmad Khan, Osamah Ibrahim Khalaf, Carlos Andrés Tavera Romero, Muhammad Sulaiman, Maharani A. Bakar, *"Application of Intelligent Paradigm through Neural Networks for Numerical Solution of Multiorder Fractional Differential Equations", Computational Intelligence and Neuroscience, vol.* 2022, Article ID 2710576, 16 pages, 2022. https://doi.org/10.1155/2022/2710576.

[15] Li, Y., Luo, F., Li, W., Zheng, S., Wu, H. H., & Xiao, C. *Self-supervised monocular depth estimation based on image texture detail enhancement.* The Visual Computer, 37(9),(2021) 2567-2580.

[16] Bhattacharyya, S., Shen, J., Welch, S., & Chen, C. *Efficient unsupervised monocular depth estimation using attention guided generative adversarial network.* Journal of Real-Time Image Processing, 18(4),(2021) 1357-1368.

[17] ZKaushik, V., Jindgar, K., & Lall, B. *ADAADepth: Adapting Data Augmentation and Attention for Self-Supervised Monocular Depth Estimation.* IEEE Robotics and Automation Letters, 6(4),(2021) 7791-7798.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: May 13, 2024
Accepted: Feb 11, 2025

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1539–1550, DOI 10.12694/scpe.v26i3.4326

A STUDY ON FAST ENGLISH SENTENCE RETRIEVAL BASED ON SIMHASH AND VECTOR SPACE MODEL TF-IDF IN AN E-LEARNING ENVIRONMENT

YUEHUA LI[*] AND XINXIN GUAN[†]

Abstract. With the rapid development of the digital information age on the Internet, information data on the Internet grows exponentially every day. In today's online learning environment, fast retrieval of English sentences plays a crucial role in the teaching and learning of modern English. The current case-based Machine translation methods can perform in-depth Parsing on sentences, and only use similar instances in the original corpus for matching and replacement processing. However, there are still certain limitations in terms of retrieval speed and similarity calculation. The study proposes an improved Simhash algorithm, which introduces substitution cost for synonym replacement and combines Term Frequency-Inverse Document Frequency (TF-IDF) weights with lexical weights for sentence-to-sentence similarity calculation. The results showed that the performance of the improved Simhash algorithm reached a maximum RI of 98.9%, an improvement of 1.4% compared to the traditional Simhash algorithm. The minimum misclassification rate of the improved algorithm was only 1.1%, a reduction of 1.4% compared to the traditional algorithm. The runtime of the improved Simhash algorithm was only 0.71s per sentence without processing synonyms and 1.82s with processing synonyms, while the runtime of the TF-IDF method alone was 71.82s and 98.11s in these two cases respectively. The improved Simhash algorithm, which combines TF-IDF weight, part of speech weight, and replacement cost, achieved an average accuracy of 92.87%, a recall rate of 88.7%, and an F1 Score of 92.87% in two calculations. This shows that the improved Simhash algorithm has high retrieval accuracy for fast retrieval of English sentences and shows excellent performance, providing a reliable technical support for the current English learning field.

Key words: Simhash algorithm; TF-IDF; Similarity; Synonym replacement; English search

1. Introduction.

The massive growth in the amount of information on the Internet has led to data overload, making it difficult for users to get exactly and quickly to the information they most want to wade through, thus increasing the cost of effort and time for users. The field of English language teaching on the Internet has also been greatly affected, and research into the rapid retrieval of English sentences has far-reaching implications for the effective implementation of English language teaching. A great deal of research has been done on sentence retrieval at home and abroad. The traditional method of sentence retrieval is to judge the similarity between sentences based on the matching degree of keywords, with more matching words indicating a higher degree of similarity. However, this method only considers individual words and appears too general [1]. Current retrieval methods divide sentence similarity into three levels: semantic, syntactic and pragmatic. However, this method is extremely difficult to implement and cannot be used in practical retrieval. Commonly used similarity detection techniques include edit distance-based, identical vocabulary-based and vector space model-based methods [2]. The identical vocabulary-based approach is relatively simple, as it only requires the number of identical words between two sentences to be judged. However, the accuracy of the calculation is lower than the other methods. The method based on vector space model is based on constructing each sentence as a high-dimensional vector and judging the semantic similarity by the cosine of the angle between the two vectors [3]. However, this method is only applicable to very few fields and cannot meet the actual large-scale and special occasion measurements, and also suffers from the problem of inaccurate calculation due to information omission. Therefore, an improved Simhash algorithm combining Term Frequency-Inverse Document Frequency (TF-IDF) weights, lexical weights and substitution costs is designed to address the above problems and applied to the fast retrieval of English sentences in order to achieve better results in sentence retrieval. The algorithm is also applied to English sentence fast retrieval with a view to achieving better application results in sentence retrieval.

[*] Basic Teaching Department, Yantai Vocational College, Yantai, 264670, China (Corresponding author, Yuehua_Li23@outlook.com)

[†] Basic Teaching Department, Yantai Vocational College, Yantai, 264670, China

2. Literature review. The proliferation of information on the Web has had a negative impact on the current English learning environment, and several researchers have conducted studies on the retrieval of English information. Fu et al. argued that the size of training data for parallel text is still limited and designed an adversarial bidirectional sentence embedding mapping structure. The structure was able to map a limited amount of parallel text data and was shown to exhibit significant advantages in low-resource environments [4]. Boban et al. proposed the use of a reverse sentence frequency method to retrieve English sentences in order to achieve a more intelligent English sentence retrieval goal and to verify the effect of different query lengths on retrieval. The results showed that the method significantly improved sentence querying [5]. Ye et al. designed an intelligent retrieval algorithm based on wireless sensor networks to address the problems of long retrieval time and low accuracy of situational English information. The algorithm uses information filtering and structured documents to achieve intelligent retrieval of English information, and the results show that the method effectively reduces retrieval time and significantly improves accuracy [6]. Kim et al. found that current visual language methods can only support up to two languages, so a modular solution was devised. The method is able to perform more language tasks by means of multimodal language embedding, and results show that it supports up to four languages with an average recall of 20.3% [7]. Khot's team designed a multi-hop inference dataset to optimise a linguistic inference model in order to achieve efficient knowledge combination from multiple texts. The model was able to retrieve and combine valid information from a large corpus of English, and the results showed that the method significantly improved retrieval performance [8].

The Simhash algorithm provides effective technical support for various fields. realizing the importance of medical knowledge graphs for the biomedical field, Wu et al. developed an inference model for reasoning about the realization of paths in combination with the Simhush algorithm and applied it to practical medical detection. The results show that the model exhibits excellent performance for medical applications [9]. Rao et al. found that visual similarity-based techniques could not detect phishing sites in legitimate regions, so Simhash with perceptual hash was introduced to calculate the similarity of phishing points and a random forest model was used to evaluate the effectiveness of the heuristic filter. The results showed that the accuracy of the model was as high as 98.73% [10]. Xiao and other researchers constructed a digital ELT hierarchical retrieval model to address the problems of low search-completeness and accuracy of traditional ELT retrieval models. The model combines the TF-IDF method with the Simhush algorithm to detect the similarity of database documents, and the results show that the model has a high completeness rate of 95% and an accuracy rate of over 96% [11]. Lin et al. considered that the current evaluation methods of network security are too complicated, and designed the Simhash model in a big data environment. The model focuses on dividing the network into multiple modules to obtain security data, and the results show that it is well adapted to large-scale data network evaluation [12].

In summary, many researchers have devised effective solutions for retrieval of English information and have achieved corresponding success in the improvement and application of the Simhash algorithm. However, few scholars have conducted experimental studies on the fusion of the two treatments, so the study introduces an improved Simhash algorithm based on the traditional Simhash algorithm and applies this to the fast retrieval of English sentences in order to obtain better practical application results.

3. Objective of the work. This article mainly explores the rapid retrieval of English sentences in the online learning environment. Traditional online English teaching has problems such as slow English information retrieval and difficulty in obtaining information, which affects students' learning effectiveness and efficiency. To solve this problem, this paper discusses the retrieval algorithm of similar cases in the case Machine translation system, and proposes a retrieval method of similar cases of English sentences suitable for large-scale corpora, in order to achieve good results in modern online English teaching. The research content mainly includes four parts. The first part mainly reviews the retrieval problem of English information and the application of Simhash algorithm. The second part mainly introduces a locally sensitive hash algorithm called Simhash, which can be used for webpage deduplication, and provides a detailed introduction to its working principle. Furthermore, the feasibility of applying it to similar case retrieval in Machine translation is discussed. At the same time, a Vector space model algorithm with high accuracy TF-IDF method is introduced. And apply the combination of the two to the rapid retrieval of English sentences. The third part verifies the retrieval effect of the proposed method on English sentences. The fourth part analyzes the experimental results to demonstrate the superiority of the proposed method. At the same time, propose areas for improvement in the research and

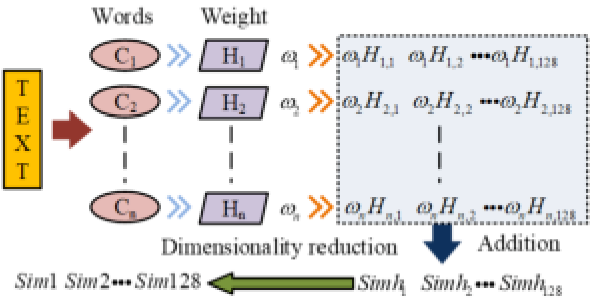


Fig. 4.1: Specific algorithm flow of Simhash

provide prospects for the future work to be done.

4. English sentence search based on TF-IDF.

4.1. Simhash algorithm based on similarity detection. The Simhash algorithm is essentially a locally sensitive hashing algorithm, which was first used in the search engine of a large number of web pages. The Simhash algorithm is used to obtain a Simhash value by dimensionality reduction of web pages, and then compare the Simhash values of different web pages using the Hemming distance to determine their similarity. The Simhash algorithm is widely used in the fields of text similarity detection, page de-duplication and sentence retrieval. Traditionally, text similarity detection is usually performed by word separation and then converted into a feature vector for distance measurement. However, the large number of feature vector words within a single text increases the dimensionality of the algorithm, which leads to an increase in computational cost and cannot be applied in larger scale environments. The Simhash algorithm aims to reduce dimensionality by mapping high-dimensional feature vectors into fixed-dimensional fingerprints through a dimensionality reduction process, and obtain the similarity of web content by comparing the fingerprints of two texts [13]. The Simhash algorithm consists of five main steps, namely word separation, hashing, weighting, merging and dimensionality reduction. The implementation steps are: firstly, the original text is divided into words to obtain the set of words$\{W_1, W_2, ..., W_n\}$ and set different levels of weights for each word in the text. The hash value of each word is then calculated using the hash, and the 0 in the hash value is turned into -1, thus transforming the set of words$\{W_1, W_2, ..., W_n\}$ into the set of$\{H_1, H_2, ..., H_n\}$, whereH_i represents the hash value of the$n$ bits. Next, the weights of each word are weighted into the$\{H_1, H_2, ..., H_n\}$ set and all the hash values in the set are accumulated in turn to obtain an bit text feature value, denoted as$\{Simh_1, Simh_2, ..., Simh_n\}$, which is calculated as shown in equation (4.1).

$$Simh_j = \sum\nolimits_{i=1}^{n} H_{ij}\mu_i \tag{4.1}$$

In equation (4.1),μ_i represents the weight of each word andH_{ij} refers to the$j$ bit of the hash value of the$i$ word. Finally, the Simhash signature is obtained by dimensionality reduction of the text feature values, which is calculated as shown in equation (4.2).

$$Sim_j = redu\left(Simh_j\right) = \begin{cases} 1 & Simh_j > 0 \\ 0 & Simh_j \leq 0 \end{cases} \tag{4.2}$$

The specific algorithm flow of Simhash is shown in Figure 4.1.

The Simhash algorithm uses the Hamming distance to determine the similarity of two pieces of data. The Hamming distance represents the number of different index positions in each of two equal strings and is calculated as shown in equation (4.3).

$$Hammin\left(x, y\right) = \sum\nolimits_{i=1}^{n} y_i \oplus x_i \tag{4.3}$$

In equation (4.3),$x = (x_1, x_2, ..., x_n)$,$y = (y_1, y_2, ..., y_n)$, and$\oplus$ represent heterogeneous operations. The Simhash algorithm converts text into signatures, which facilitates retrieval and also plays a space-saving role. The similarity of two texts can be calculated by the Hemming distance of the signature, as shown in equation (4.4).

$$sim\left(T_1, T_2\right) = \frac{\sum_{k=1}^{128} T_{2k} \oplus T_{1k}}{128} \tag{4.4}$$

In equation (4.4),T_{1k} andT_{2k} refer to the value at the$k$ bit of the two signatures and 128 represents the number of bits in the string. The smaller the Hemming distance of the signatures, the higher the similarity of the two texts. The traditional Simhash algorithm mainly uses the number of occurrences of feature terms as the weight of the weighting, which will result in the Simhash signature not accurately characterising the textual information [14]. Therefore, the study introduces the TF-IDF value and combines it with the lexical properties of words to jointly calculate the weights of feature items. TF-IDF is a numerical weighting calculation method commonly used in natural language processing, which is widely used in information retrieval, text clustering and other fields. The TF refers to word frequency, which represents the number of times a word appears in a text, and IDF means inverse text frequency index, which characterises the text differentiation ability of a word. The weight of a word is calculated as the product of TF and IDF [15-16]. The TF value is calculated as shown in equation (4.5).

$$TF_{ij} = \frac{b_{ij}}{\sum_l b_{ij}} \tag{4.5}$$

In equation (4.5),b_{ij} represents the number of occurrences of the$i$ feature word of the$j$ text in the text, and$l$ refers to the set of all words in the text. The IDF value is calculated as shown in equation (4.6).

$$IDF_{ij} = \log \frac{|D|}{|\{d : t_{ij} \in d\}|} \tag{4.6}$$

In equation (4.6),D represents the set of texts, andt_{ij} represents the$i$ th feature word of the$j$ th text. Let the total number of data in the text dataset be N, and the feature wordt_{ij} exists in the$c_{i,j}$ text, the IDF value is calculated as shown in equation (4.7).

$$IDF_{ij} = \log \frac{N}{\beta + c_{i,j}} \tag{4.7}$$

In equation (4.7),β is generally taken as 1, which serves to prevent the denominator from being 0. The weights of the feature terms are calculated as shown in equation (4.8).

$$w_{ij} = IDF_{i,j} \bullet TF_{i,j} \tag{4.8}$$

Assume two text vectors as shown in equation (4.9).

$$\begin{aligned} &\{W_1 - T_1, W_2 - T_2, ..., W_N - T_N\} \\ &\{W_1 - T_1', W_2 - T_2', ..., W_N - T_N'\} \end{aligned} \tag{4.9}$$

The similarity of two texts can be determined by the cosine of the angle between the two vectors, calculated as shown in equation (4.10).

$$Similarity(V, V') = \frac{\sum_{i=1}^{n} T_i \bullet T_i'}{\sqrt{\sum_{i=1}^{n} {T_i}^2 \bullet \sum_{i=1}^{n} T_i'^2}} \tag{4.10}$$

In equation (4.10),V andV' represent two vectors. TF-IDF is feasible for the calculation of feature item weights and text similarity, but it only considers the number of occurrences of feature items, ignoring the influence of different lexicalities on the semantic expression of the text. Therefore, the study uses the lexicality

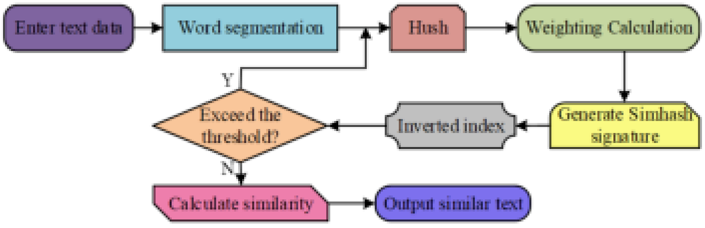


Fig. 4.2: Specific process of improved Simhash algorithm

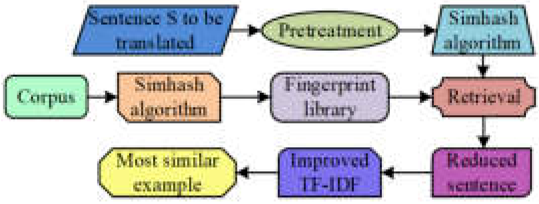


Fig. 4.3: Flow chart of fast English sentence retrieval

of feature words as a measure of word weights as well. The final feature word weights are calculated as shown in equation (4.11).

$$w_{ij} = w_k \bullet IDF_{i,j} \bullet TF_{i,j} \tag{4.11}$$

In Eq. (4.11),w_k refers to the weights corresponding to the specified lexical properties. The noun weight is 4, the verb is 3, the adjective is 2 and the rest of the lexical nature is 1. The combination of TF-IDF value and lexical nature is introduced into the feature weight calculation, which can characterise the text content more comprehensively, thus increasing the effectiveness of the Simhash algorithm for text similarity detection. The specific flow of the improved Simhash algorithm is shown in Figure 4.2.

4.2. Simhash-based Fast English Sentence Retrieval. The study is based on the Simhash algorithm and the vector space model TF-IDF for fast retrieval of English sentences. Among them, the Simhash algorithm focuses on quickly detecting similar texts from a large amount of information data and then returning the text set [17]. The vector space model-based TF-IDF method, on the other hand, focuses more on the accurate representation of the internal information of the text. In a practical translation system, the sentence with the highest similarity to the sentence to be translated needs to be retrieved quickly from a large-scale corpus. Therefore, the study combines the Simhash algorithm with the TF-IDF method to design an algorithm for retrieving the most similar text instances with high accuracy. The algorithm first selects sentences with high similarity by generating a fingerprint library through the Simhash algorithm to form a reduced set of sentence instances. Then a synonym dictionary is applied to all sentences and the replacement cost is calculated. Finally, the improved TF-IDF is used to construct the feature vector and calculate the similarity between each sentence and the sentence to be translated, so as to find the example sentence with the maximum similarity to the sentence to be translated. The flow of the algorithm is shown in Figure 3.

In the Simhash algorithm, the Hemming distance between fingerprints is calculated by first using a heterogeneous operation and then checking the number of ones in the result. Google's web de-duplication algorithm uses the drawer principle to create an inverted index to calculate the Hemming distance. The study refers to this algorithm to calculate the Hemming distance by grouping fingerprints for retrieval. For a 32-bit fingerprint, the threshold is set to 7 and the fingerprint is divided into 8 equal parts of 4 bits each. The specific operation flow is shown in Figure 4.4.

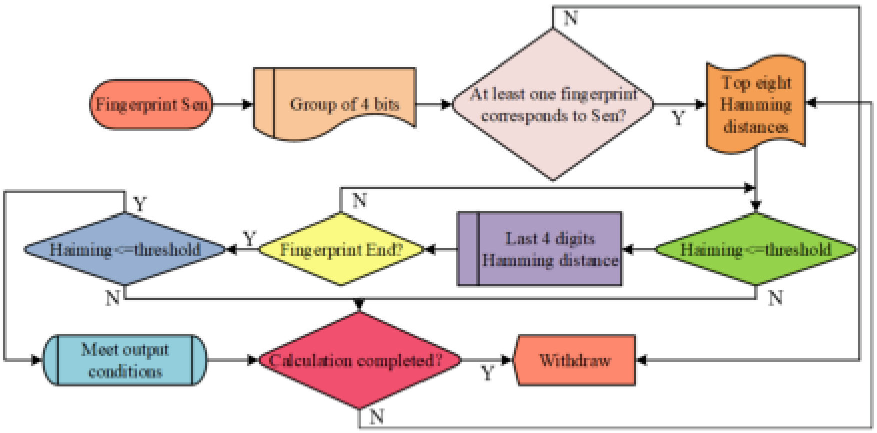


Fig. 4.4: Calculation of Hamming Distance of Fingerprint

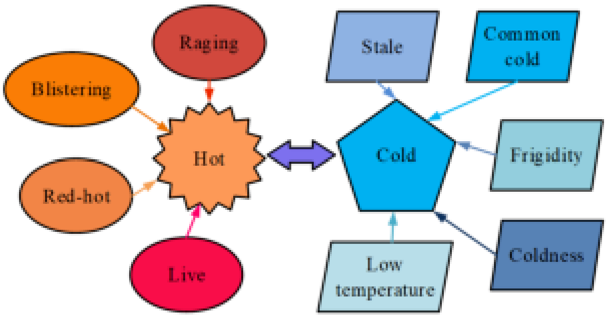


Fig. 4.5: Synonyms of hot and cold

After narrowing down the range of similar sentences using the Simhash algorithm, the similarity of each instance to the sentence to be translated was then calculated. However, the traditional vector method is unable to identify semantic information, resulting in low similarity calculation results. Therefore, the study uses the sentence to be translated, S, as a benchmark and replaces all words that have a synonymous relationship with S with words that exist in [18-19]. This method can make the similarity calculation results more accurate, and at the same time has a good effect of dimensionality reduction, which in turn simplifies the calculation of TF-IDF. However, since there are multiple meanings in natural language, direct substitution after detecting synonyms will lead to substitution errors. For this reason, the study further introduces a parameter to measure the correct rate of substitution between synonyms, namely the substitution cost. When the substitution is correct, the substitution cost is 1, which means that the substitution is possible; when the substitution is incorrect, the substitution cost is 0, and the substitution should not be made. When the replacement is not necessarily correct, then the replacement cost is between 0 and 1. The replacement cost allows the weight of the replaced position to be reduced, thus optimising the results of the calculation. WordNet is currently used to process English vocabulary, not only for lexical purposes but also for semantic word relations, including antonymy, synonymy, subordination and whole-part relations. Antonymic relations are usually found between adjectives, which characterise semantic information through an N-dimensional hyperspace structure, and use antonymic relations to link different clusters of synonyms. For example, the synonymic clusters of hot and cold are represented as shown in Figure 4.5.

When two words are substituted for each other in a linguistic text without affecting the original semantics,

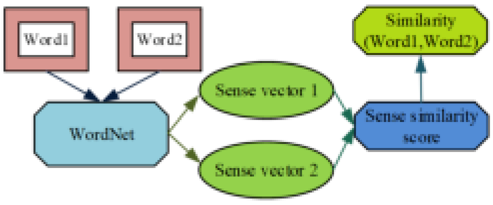


Fig. 4.6: Calculation process of word meaning similarity

the two words are in a synonymic relationship. The superior-subordinate relationship is also referred to as the parent-child relationship or ISA relationship, which is transitive in nature. A whole-part relationship is one in which the meaning of a word is part of another word set [20]. The substitution cost is mainly used to calculate the correctness of the substitution operation, and the lexical similarity of the two words is usually used as the substitution surrogate value. The process of calculating the lexical similarity is shown in Figure 4.6.

When calculating word sense similarity using WordNet, feature extraction of the word sense is first required. The extraction is calculated as shown in equation (4.12).

$$Feature(SW) = \{\{WE\},\{WC\},\{WS\}\} \tag{4.12}$$

In equation (4.12),WE represents all real words in the interpretation of$W$,WC refers to all related genera, andWS represents all synonyms of$W$ in WordNet. The similarity of two words can be obtained by calculating their distances in different feature spaces; the further the distance, the smaller the similarity. The similarity is calculated as shown in equation (4.13).

$$Similarity(SW_i, SW_j) = \frac{1}{No(SW)\bullet No(SW_j)} \times \frac{\sum_{W_i \in \{W_{Si}\} \cap \left\{W_{Sj}\right\}} IDF(w_i)^2 \bullet K_S + \sum_{W_i \in \{W_{Ci}\} \cap \left\{W_{Cj}\right\}} IDF(w_i)^2 \bullet K_C + \sum_{W_i \in \{W_{Ei}\} \cap \left\{W_{Ej}\right\}} IDF(w_i)^2 \bullet K_E}{\sqrt{\sum_{i \in Q_o, K \in \left\{K_E, K_C, K_S\right\}} IDF(w_i)^2 \bullet K \times \sum_{j \in Q_p, K \in \left\{K_E, K_C, K_S\right\}} IDF(w_j)^2 \bullet K}} \tag{4.13}$$

In equation (4.13),$IDF\,(w_i)$ represents the countdown of the number of times a text appears when WordNet is created. $No(SW)$represents the order of meaning of the words. $K_E K_C$ andK_S represent the feature weights of sense interpretation, class attributes and synonyms respectively. $Q_O Q_P$ refers to the set of indicators wherew_i appears;w_j refers to the set of indicators where and appear. IfSW_1 andSW_2 are used to denote the number of senses ofW_1 andW_2 , respectively, the word sense similarity is calculated as shown in equation (4.14).

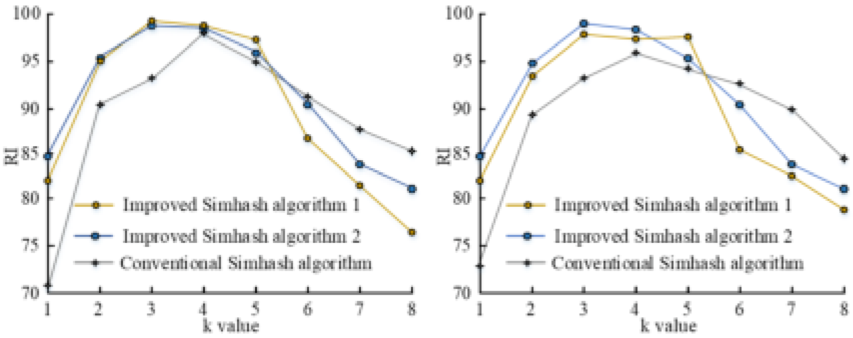
$$Similarity(W_1, W_2) = \frac{\sum_{i \in \{1,\ldots,|SW1|\}, j \in \{1,\ldots,|SW2|\}} S(SW1_i, SW2_j)}{|SW1| + |SW2|} + \frac{\sum_{i \in \{1,\ldots,|SW2|\}, j \in \{1,\ldots,|SW1|\}} S(SW2_i, SW1_j)}{|SW1| + |SW2|} \tag{4.14}$$

The research further incorporates the replacement cost into the TF-IDF algorithm to improve the algorithm. The steps of the improved TF-IDF algorithm are as follows: first, WordNet is used to compare the translated sentence S with each example sentence in the narrowed down instance library E, and synonym pairs are found. Calculate the semantic similarity α of the synonym pair, and use this as the substitution value of the synonym pair. Next, replace synonyms with the words appearing in the sentence S to be translated, multiply by α at the replacement position, and construct feature vectors for E and S. And finally the similarity between the two sentences is calculated and the most similar instance S' is obtained, whose similarity is calculated as shown in equation (4.15).

$$Similarity = \frac{\sum_{i=1}^{n} \omega_i \bullet \omega_i'[\alpha]}{\sqrt{\sum_{i=1}^{n} (\omega_i[\alpha])^2} \times \sqrt{\sum_{i=1}^{n} (\omega_i')^2}} \tag{4.15}$$

Table 5.1: Experimental environment and configuration

Operate	Ubuntu 16.04
Memory	128G
Hard disk	10T
Programming language	Python3.6
GPU	GTX 1080
Deep learning framework	TensorFlow
CPU	Intel Xeon E5-2682 v4



(a) RI value of the first test

(b) RI value of the second test

Fig. 5.1: RI values for different algorithms

5. Experimental analysis of English sentence retrieval based on Simhash algorithm. All experiments were conducted on the server host in the laboratory, using TensorFlow, a deep learning framework developed by Google for data flow programming for multiple tasks, as the implementation tool for the network model. Accelerate training with a single NVIDIA GTX 1080 graphics card. The experimental environment and configuration are shown in Table 5.1.

In order to verify the effectiveness of the improved Simhash algorithm for retrieval of English sentences, the study first conducted performance tests on the clustering effect of the algorithm and chose the traditional Simhash algorithm to compare the results. Due to the fact that online news conforms to the density connected model, the test data was selected from a 200W scale English news dataset, and 400 of them were randomly selected for evaluation experiments.The study selected the RI (Rand Index) value as the performance testing indicator, and the larger the RI value, the better the clustering effect. In general, an RI value of 90% is required to meet practical application requirements. Since the number of segments k has a large impact on the clustering effect, different k values need to be set, and the experiment was conducted twice.

In Figure 5.1, the same clustering evaluation method and manual discriminant method as the traditional Simhash algorithm were used for the improved Simhash algorithm respectively, resulting in two different curves. From the results of the two tests, it can be seen that the improved Simhash algorithm has the highest RI value when the value of k is taken as 3, and it reaches a maximum of 98.9%. When the value of k is greater than 3, the clustering effect will then decrease. The reason for this is that too large a value of k will cause otherwise unrelated clusters to be combined together, thus reducing the clustering accuracy. The traditional Simhash algorithm takes the highest RI value of 97.5% when k is taken to 4, a decrease of 1.4% compared to the improved Simhash algorithm. The misclassification rate is equal to 1 - the RI value, and the minimum misclassification rate of the traditional Simhash algorithm is 2.5%, while the minimum misclassification rate of the improved algorithm is only 1.1%.

Further evaluate the effectiveness of the improved Simhash algorithm in fast English sentence retrieval, using three indicators: accuracy, recall, and F1 Score to measure the algorithm's effectiveness. Among these

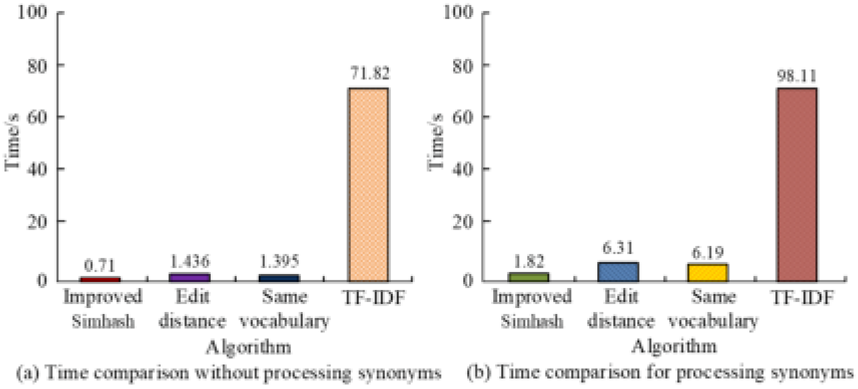


Fig. 5.2: Comparison of running times of different methods

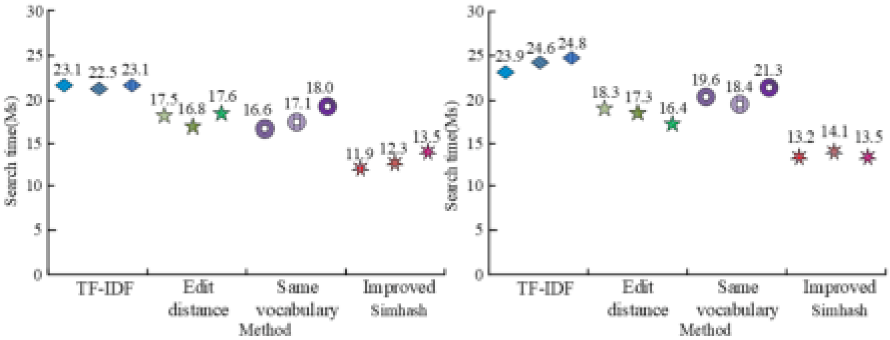
three methods, the edit distance and identical vocabulary-based methods only consider the identical words in two sentences, ignoring the corpus as a whole, which results in lower computational accuracy. The vector-based TF-IDF method, on the other hand, considers different words, identical words and the influence of each word in the corpus on the sentence, and is therefore more accurate. Thirty English sentences from the corpus were used to compare the running time of the four algorithms. was used to record the running time of each sentence instance and to calculate the average time spent by each algorithm. At the same time, the English-Chinese parallel corpus was chosen as the experimental corpus, with a size of 9,948 pairs. The experiments were conducted separately to compare and analyse the case of no synonym processing with the case of introducing synonym processing, and the experimental results are shown in Figure 5.2 8.

As can be seen from Figure 5.2 , the TF-IDF algorithm alone has the longest running time for each sentence, 71.82s and 98.11s respectively, both in the case of no synonyms and in the case of synonyms. the reason for this is that the TF-IDF algorithm requires a high-dimensional vector construction for each sentence instance, which leads to a significant time consumption. The experimental results based on the same vocabulary and edit distance methods are not very different, around 1.4s versus 6.2s in the two cases respectively. In contrast, the improved Simhash algorithm proposed in the study runs in only 0.71s and 1.82s in both cases, which can be seen to have a significant advantage in terms of time performance. In addition, the TF-IDF method has a larger increase in runtime after the introduction of synonym processing compared to the other three methods, mainly because it not only performs synonym queries but also calculates word similarity. In contrast, the method based on edit distance and identical words only performs synonym queries. The improved Simhash algorithm reduces the time consumption for synonym processing as the range of similar instances is reduced, thus reducing the number of queries for synonyms and calculating word similarity. The study continued to measure the retrieval time for different sizes of text fingerprints and repeated the measurement three times to take the average value.

As can be seen from Figure 5.3, the retrieval speed of fingerprints slows down as the size of the fingerprint library increases. The improved Simhash algorithm also has the lowest average retrieval time of 12.6ms and 13.6ms for fingerprints of size 180,000 and 200,000 respectively, which is 10.3ms and 10.8ms respectively compared to the TF-IDF method.

Since the TF-IDF method considers factors such as different words, identical words and the influence of each word on the sentence at the same time, its computational accuracy is relatively high. Therefore, experiments were conducted to analyse its similarity optimisation results with the improved Simhash algorithm. When synonyms were correctly replaced, some of the test results are shown in Table 5.2.

In Table 5.2, a represents the experimental input sentence and b represents the output sentence after synonymous replacement. As can be seen from Table 1, when the synonyms in the sentence instances are correctly replaced, then the Simhash algorithm proposed by the study computes higher similarity results than the TF-IDF method alone. In the tested sentence pairs, the Simhash algorithm achieves a maximum similarity result of 0.9542, which is close to the true value and 0.1217 higher than that of the TF-IDF method, while the



(a) Retrieval time when fingerprint size is 180000

(b) Retrieval time when fingerprint size is 200000

Fig. 5.3: Retrieval time under different scale fingerprints

Table 5.2: Similarity Results of Correct Substitution of Synonyms

Test sentence	Improved Simhash	TF-IDF
a: Word can not depict the wonders of nature	0.8521	0.5075
b: Word can't describe the beauty of the scene		
a: She believes that the present continent once including his name after Ultima of Pangea	0.8369	0.7214
b: She once proposed that the present mainland including a continent he named Pangaea		
a: Please recommend a shoe store to me	0.9542	0.8325
b: I am looking for a less expensive store		

Table 5.3: Similarity results of synonyms being incorrectly replaced

Test sentence	TF-IDF	Improved Simhash
a: I begged Betty to give me some staples	0.5367	03465
b: Betty begged Mary to take a course for him		
a: This is a brightly lit room with high windows	0.2443	0.0136
b: This paper uses substantive cases to popularize it and transparent		
a: The teacher recorded my grades on the form	0.3798	0.0498
b: I will reserve a table for eight		

difference in similarity reaches a maximum of 0.3446. The result will be higher than the true value. Some of the results of the tests when synonyms were replaced incorrectly are shown in Table 5.3.

As can be seen from Table 3, the improved Simhash algorithm, which introduces the cost of synonym substitution, computes significantly lower similarity results when synonyms are replaced incorrectly. The reason for this is that the algorithm reduces the adverse effect of incorrect substitution on the selection of similar sentences and effectively optimises the calculation of similarity. The degree of optimisation mainly depends on the size of the replacement generation value, the smaller the replacement cost the greater the probability of incorrect replacement occurring, thus leading to an improved optimization effect; the larger the replacement cost the smaller the probability of incorrect replacement occurring, then the optimisation effect is not significant.

The study further evaluates the effectiveness of the improved Simhash algorithm in English sentence fast retrieval, using three metrics: accuracy, recall and F1-Score to measure the effectiveness of the algorithm. At the same time, two traditional Simhash algorithms were chosen for comparison experiments with the improved

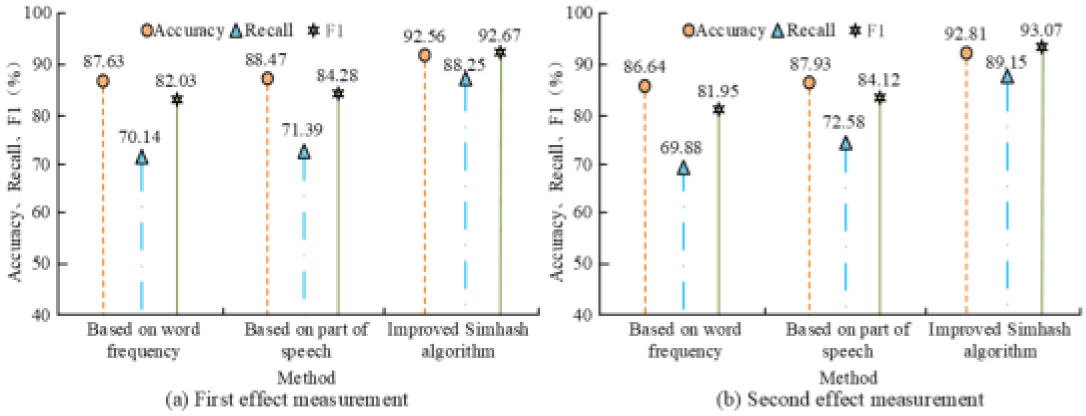


Fig. 5.4: Accuracy, recall, and F1-Score of different algorithms

Simhash algorithm, including the word frequency-based Simhash algorithm and the lexicality-based Simhash algorithm. The results of the two measurements are shown in Figure 5.4.

As can be seen from Figure 5.4, the improved Simhash algorithm that combines TF-IDF weights, lexical weights and replacement cost has an average accuracy of 92.87%, recall of 88.7% and F1-Score of 92.87% for the two measurements. Compared with the traditional Simhash based on word frequency weights, the improvement was 5.735%, 18.69% and 10.88% respectively. It indicates that the fusion of TF-IDF weights, lexical weights and substitution cost can distinguish the influence of different feature words on the text, thus enabling more feature information to be included in the obtained text fingerprint. At the same time, the improved Simhash algorithm effectively enriches the semantic information of the lexicon, which in turn significantly improves the correct substitution rate between synonyms and optimises the similarity calculation results to make them closer to the true value.

6. Conclusion. The complex web-based learning environment has made fast English sentence retrieval an important method for modern English learning. The study combines the TF-IDF method to develop an improved Simhash algorithm, which introduces substitution cost for synonym processing of English sentences and integrates TF-IDF weights with lexical weights for similarity calculation. The results show that the similarity values calculated by the improved Simhash algorithm are very close to the true values when the synonyms are correctly replaced. At the same time, when synonyms were replaced incorrectly, the similarity values calculated by the improved Simhash algorithm were not inflated. The lowest calculated value is 0.0136, which is 0.2307 lower compared to the TF-IDF method. meanwhile, the average retrieval time of the improved Simhash algorithm is 12.6ms and 13.6ms for fingerprints of 180,000 and 200,000 scale respectively. it is 10.3ms and 10.8ms lower compared to the TF-IDF method respectively. in addition, the improved the accuracy, recall and F1-Score of the Simhash algorithm effect reached 92.87%, 88.7% and 92.87%, respectively. Compared with the Simhash based on word frequency weights, they rose by 5.735%, 18.69% and 10.88% respectively. It indicates that the improved Simhash algorithm is highly feasible for fast retrieval of English sentences and has excellent performance qualities.This method can effectively solve the problems of slow retrieval speed and low accuracy in current online English teaching. It effectively improves the learning efficiency and effect of students, optimizes the online teaching effect of Modern English, and promotes the development of modern online English teaching.but there is still some room for improvement in its retrieval accuracy, and thus further improvement is needed.

Acknowledgement. The research is supported by Yantai Vocational Education Research Project in 2023: ”A Study on the Digital Competence of English Teachers in Higher Vocational Colleges under the Background of Educational Digital Transformation” (Higher Vocational Colleges 2023-5); Yantai Vocational College Horizontal Scientific Research Project: ”Development of the English Service Capacity Improvement Project for Yantai

Microspring Youth Affairs Service Center" (HX2024073).

REFERENCES

[1] J. Qin, *An Encrypted Image Retrieval Method Based on SimHash in Cloud Computing*, Computers, Materials and Continua, vol. 62, no. 3, pp. 389–399, 2020.

[2] R. H. Dong, C. Shu, Q. Y. Zhang, *Security Situation Assessment Algorithm for Industrial Control Network Nodes Based on Improved Text SimHash*, International Journal of Network Security, vol. 23, no. 6, pp. 973–984, 2021.

[3] M. J. Lim, Y. M. Kwon, *Efficient algorithm for malware classification: N-gram MCSC*, International Journal of Computing and Digital Systems, vol. 9, no. 2, pp. 179–185, 2020.

[4] Z. Fu, Y. Xian, S. Geng, Y. Ge, Y. Wang, X. Dong, G. De Melo, *ABSent: Cross-lingual sentence representation mapping with bidirectional GANs*, Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 5, pp. 7756–7763, 2020.

[5] I. Boban, A. Doko, S. Gotovac, *Sentence retrieval using stemming and lemmatization with different length of the queries*, Advances in Science, Technology and Engineering Systems, vol. 5, no. 3, pp. 349–354, 2020.

[6] Q. Ye, *Situational English Language Information Intelligent Retrieval Algorithm Based on Wireless Sensor Network*, International Journal of Wireless Information Networks, vol. 28, no. 3, pp. 287–296, 2021.

[7] D. Kim, K. Saito, K. Saenko, S. Sclaroff, B. Plummer, *Mule: Multimodal universal language embedding*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 11254–11261, 2020.

[8] T. Khot, P. Clark, M. Guerquin, P. Jansen, A. Sabharwal, *Qasc: A dataset for question answering via sentence composition*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 5, pp. 8082–8090, 2020.

[9] X. Wu, J. Duan, Y. Pan, M. Li, *Medical knowledge graph: Data sources, construction, reasoning, and applications*, Big Data Mining and Analytics, vol. 6, no. 2, pp. 201–217, 2023.

[10] R. S. Rao, A. R. Pais, *Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach*, Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 9, pp. 3853–3872, 2020.

[11] Z. Xiao, *Ontology-based hierarchical retrieval model for digital English teaching information*, International Journal of Continuing Engineering Education and Life Long Learning, vol. 33, no. 2-3, pp. 337–350, 2023.

[12] P. Lin, Y. Chen, *Network Security Situation Assessment Based on Text SimHash in Big Data Environment*, International Journal of Network Security, vol. 21, no. 4, pp. 699–708, 2019.

[13] J. Qin, *An encrypted image retrieval method based on SimHash in cloud computing*, Computers, Materials & Continua, vol. 63, no. 1, pp. 389–399, 2020.

[14] S. Fedushko, *Scientific Content: Language Expansion in Bibliometric Databases*, 2020.

[15] S. Tang, *Identification of Scratch projects' Similarity Using Clustering Algorithms*, International Core Journal of Engineering, vol. 7, no. 12, pp. 158–170, 2021.

[16] S. Fedushko, O. Trach, Z. Kunch, Y. Turchyn, U. Yarka, *Modelling the behavior classification of social news aggregations users*, arXiv preprint arXiv:1909.01677, 2019.

[17] Q. Ye, *RETRACTED ARTICLE: Situational English Language Information Intelligent Retrieval Algorithm Based on Wireless Sensor Network*, International Journal of Wireless Information Networks, vol. 28, no. 3, pp. 287–296, 2021.

[18] R. S. Rao, A. R. Pais, *Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach*, Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 9, pp. 3853–3872, 2020.

[19] Z. Xiao, *Ontology-based hierarchical retrieval model for digital English teaching information*, International Journal of Continuing Engineering Education and Life Long Learning, vol. 33, no. 2-3, pp. 337–350, 2023.

[20] X. Zhang, P. Li, X. Ma, Y. Liu, *Railway wagon flow routing locus pattern intelligent recognition algorithm based on SST*, Smart and Resilient Transport, vol. 2, no. 1, pp. 3–21, 2020.

Edited by: Sheikh Nasrullah
Special issue on: Transformative Horizons:
The Role of AI and Computers in Shaping Future Trends of Education
Received: Jul 22, 2024
Accepted: Aug 2, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1551–1558, DOI 10.12694/scpe.v26i3.4336

EDUCATIONAL DATA MINING FOR STUDENT PERFORMANCE PREDICTION

LINQIANG TANG[*] AND CHEN SIAN[†]

Abstract. The topic of Educational Data Mining (EDM) has gained significant traction in improving the quality of education by identifying patterns and insights through the analysis of data gathered from diverse educational settings. In order to discover important elements that affect educational achievement and to give educators and policymakers with useful insights, this study investigates the use of machine learning techniques in predicting student performance. We use a variety of machine learning methods, such as decision trees, support vector machines, and neural networks, to create predictive models by utilizing past educational information, demographics, and behavioral tendencies. The study assesses these models' efficacy and accuracy while also emphasizing how important choosing features and data preparation are to enhancing prediction results. Our results show that applying machine learning approaches can greatly improve the prediction of pupil achievement, which in turn allows for more focused interventions and individualized learning plans. This study highlights the possibilities of machine learning in promoting a data-driven method to educational improvement and adds to the expanding body of knowledge in EDM.

Key words: Educational Data Mining, student performance, prediction, machine learning methods, Educational.

1. Introduction.

The abundance of data available in today's educational environment has made it possible for creative methods to improve student learning outcomes [8]. The discipline of Educational Data Mining (EDM) is gaining importance as it uses advanced data mining methods to examine educational data and derive important insights. EDM looks for patterns and trends in the vast amounts of data collected by educational institutions in order to anticipate student performance [10]. This allows teachers to customize interventions and methods for the best possible learning outcomes. EDM is important because it can convert unprocessed data into useful knowledge. Big data and advanced analytics have made it possible for educators and academics to have a deeper knowledge of the many variables influencing students' achievement [15].

A wide range of factors, including behavioral data, socioeconomic indicators, academic records, and demographic information, are included in EDM and contribute to a thorough knowledge of how students perform [6]. A fundamental component of EDM is predictive modeling, which makes use of statistical methods and algorithms to project future results from past data. These algorithms have the ability to pinpoint kids who are at danger, forecast grades, and even provide individualized learning plans [16]. Through the utilization of artificial intelligence and machine learning, EDM not only improves the precision of predictions but also offers insights into the fundamental elements influencing the achievement of students [18].

The increasing awareness of data-driven approaches' potential to transform teaching methods is motivated this research. Large volumes of data are produced when educational institutions use digital tools and systems more frequently, recording numerous facets of student behavior, academic achievement, and demographic traits. Even with this wealth of data at their disposal, many educational institutions nevertheless depend on antiquated, one-size-fits-all strategies that underutilize the insights this data may provide for enhancing student results.

In order to forecast and evaluate student performance, Educational Data Mining (EDM) for Student Performance Prediction applies data mining techniques to educational data. With the goal of assisting educators, administrators, and policymakers in enhancing the educational process and results, EDM seeks to reveal hidden patterns, trends, and insights from educational data [13]. The purpose of this introduction is to highlight the revolutionary possibilities of Educational Data Mining in the field of predicting student performance. Institutions can cultivate a data-driven culture that encourages academic success, reduces dropout rates, and supports students' holistic development by methodically examining educational data [17]. As we learn more about this

[*] Zhejiang Institute of Communications, Hangzhou, Zhejiang, 311112, China (tanglinqianglearni@outlook.com)
[†] Zhejiang Institute of Communications, Hangzhou, Zhejiang, 311112, China.

area, it becomes clearer how EDM has the potential to completely transform education and provide hope for a better educated and functioning educational system. The main contribution of proposed method is given below:

1. This research's primary contribution to Educational Data Mining (EDM) for student performance prediction is the creation of a solid, data-driven framework that makes use of cutting-edge machine learning algorithms to predict educational results with high accuracy.
2. In order to produce a comprehensive prediction model, this study takes into account a variety of behavioral, demographic, and socioeconomic characteristics in addition to standard academic indicators.
3. Through a methodical examination of extensive datasets, the study pinpoints significant trends and indicators of student achievement, providing educators and decision-makers with valuable perspectives.
4. This work stands out for its innovative integration of multifaceted data sources and state-of-the-art analytical approaches, which has a substantial positive impact on the development of analytics for prediction in education.

The rest of our research article is written as follows: Section 2 discusses the related work on various educational data Mining. Section 3 shows the algorithmprocess and general working methodology of proposed work. Section 4 evaluates theimplementation and results of the proposed method. Section 5 concludes the work anddiscusses the result evaluation.

2. Related Works. Education data mining (EDM) is a rapidly developing field that analyzes data from educational environments with the goal of improving education. Predicting student performance has become a major area of study for EDM because of its potential to enhance educational results [5]. Highlighting numerous strategies, models, and conclusions, this section examines significant contributions and methodology in this field. Research have used a variety of data sources such as educational records, social media activity, interaction logs from Learning Management Systems (LMS), and student demographic data [3, 20]. The author, for example, used LMS interaction data in conjunction with demographic information and past academic performance to forecast future performance.

For forecasting algorithms to be reliable and efficient, efficient techniques for preprocessing including choosing features, data cleaning, and normalization are essential. Numerous machine learning techniques have been used to forecast student achievement [9]. Diverse degrees of success have been shown by decision trees, artificial neural networks, and ensemble techniques like random forests [2] Models based on deep which can identify intricate patterns in big datasets, seem to be the direction of recent advancements. By comparing multiple algorithms, for instance, the author came to the conclusion that ensemble approaches perform better than single classifiers in most cases [11].

An important factor in the effectiveness of models for prediction is feature design. Numerous aspects have been investigated by researchers, such as involvement in online forums, assignment submission deadlines, attendance records, and even psychological elements like ambition and self-control [14]. Principal component analysis and correlation-based choice of features are two feature selection strategies that have been used to improve the performance of models by minimizing over fitting and complexity. Many metrics, such as accuracy, recall, F1 score, precision, and Area Under the Receiver Operating Characteristics Curve (AUC-ROC), are used to assess the efficacy of models for prediction. The significance of false positives vs false negatives and the particular educational setting are two factors that frequently influence the selection of metrics [4, 1].

In EDM, models of prediction are frequently used in conjunction with strategies for intervention meant to raise student achievement. Predictive analytics has been used to create tailored feedback, early warning systems, and adaptable learning environments [19]. For instance, the author developed a system for early detection that, by giving at-risk students targeted help, greatly increased student retention rates. Predictive model use in education brings up a number of ethical issues, mostly with regard to algorithmic bias, informed consent, and data protection [7, 12]. In order to guarantee responsible utilization of student data, the author talked on the significance of open data practices and the requirement for ethical principles. They support involving all relevant parties in the creation and application of predictive structures, such as teachers, pupils, and legislators.

This problem can be effectively solved with the help of Educational Data Mining (EDM), which makes it possible to analyze educational data systematically and find links and patterns that can guide decision-making.

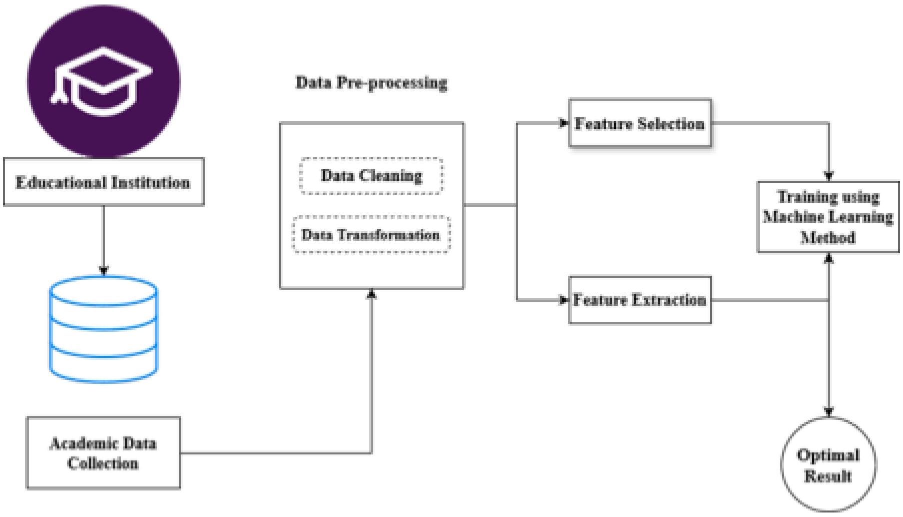


Fig. 3.1: Architecture of Proposed Method

One major development in this area is the use of machine learning algorithms to predict student performance. Educators and policymakers can improve the quality of education by implementing targeted interventions that are tailored to the specific needs of children by precisely forecasting which kids are at danger of underperforming.

3. Proposed Methodology. The proposed methodology is to enhance academic results and comprehend students' learning processes, educational information mining, or EDM, applies methods of data mining to educational data. This methodology's main goal is to forecast pupil achievement utilizing a variety of methods for data mining and instructional information. Using academic data analysis tools, the suggested methodology offers an extensive structure for forecasting student performance. Educational organizations can apply tactics to improve educational results and acquire helpful insights into pupil retention processes by adhering to this methodology. In figure 3.1 shows the architecture of proposed method.

Academic data is first gathered from the educational institution and pre-processed to remove discrepancies and convert the data into a format that may be used. To improve representation, the most pertinent features are found and additional features are created through feature extraction and selection. An ideal predictive model for student performance is produced by training a machine learning model using these features. To make sure this model is accurate and reliable, its performance is assessed using a variety of measures.

3.1. Data Collection. The original source of academic information, such as attendance, performance, and other pertinent data records for students. Offers unprocessed academic data for study. combines information from many educational institution sources. a thorough database with all the pertinent student data.

3.2. Data Pre-processing. Pre-processing data entails Managing missing values: removing or imputation, Eliminating duplicates, fixing mistakes in data entry. Data analysis outcomes can be distorted by irrelevant or meaningless information, which is referred to as noise in the data. Noise can originate from a number of things, including data entry mistakes, malfunctioning sensors, and anomalies that do not accurately reflect the dataset. Methods including filtering, outlier detection, and smoothing are used to get rid of noise. Predictive models become more accurate as noise is reduced because the data is more representative of the real underlying patterns.

3.2.1. Data Cleaning. Entails eliminating noise, dealing with missing data, and fixing inconsistent results.

$$Cleaned\ Data = Raw\ data - Noise \tag{3.1}$$

3.2.2. Data Transformation. In educational datasets, missing data is a prevalent problem. Records may be incomplete for a number of reasons, including student absences, incomplete grades, or mistakes in data collecting. In order to handle missing data, records with large gaps must be removed, or missing values must be imputed using statistical techniques.

$$Transformed\ Data = \frac{cleaned\ data - \mu}{\sigma} \tag{3.2}$$

Ready-to-use preprocessed data for the extraction and selection of features.

The process of transforming cleaned data into a format appropriate for analysis and modeling is known as data transformation. Making sure the data is consistent and suitable for machine learning algorithms requires taking this crucial step. Rescaling data to have a mean ( ) of zero and a standard deviation ( ) of one is the process of standardization. This approach guarantees that every feature contributes equally to the model, which is especially crucial for algorithms that rely on distance measurements, such neural networks or support vector machines. By addressing variables that may have varying scales, standardization helps keep characteristics with larger scales from unduly affecting the model.

3.3. Feature Selection. Determines the most important characteristics that go into predicting a student's achievement. methods such as feature importance from models, mutual information, and correlation analysis.

$$Selected\ Features = \mathrm{arg}max_{Fi}\ \ Importance\ (F_i) \tag{3.3}$$

3.4. Feature Extraction. Uses the available data to generate new features that more accurately capture the underlying patterns.

3.4.1. Principal Component Analysis (PCA). In order to simplify complex datasets and preserve as much variability (information) as possible, data analysts employ principal component analysis (PCA), a dimensionality reduction approach. Principal component analysis (PCA) aims to convert the original data into a new, uncorrelated set of features known as principal components. The arrangement of these elements ensures that the majority of the variety found in the original dataset is retained in the first few.

Scaling the characteristics to have a mean of 0 and a standard deviation of 1 is the process of standardizing the data.

$$Z = \frac{X - \mu}{\sigma} \tag{3.4}$$

The degree of collective feature variation is measured by the covariance matrix. The covariance matrix for a dataset with n features is an n×n matrix.

$$\Sigma = \frac{1}{n-1} Z^T Z \tag{3.5}$$

The covariance matrix's eigenvalues and eigenvectors are calculated. The new feature space's direction is determined by the eigenvectors, while its magnitude, or relevance, is determined by the eigenvalues.

$$\Sigma v = \lambda v \tag{3.6}$$

The eigenvalues have been arranged in descending order by their corresponding eigenvectors. The first principal component is the eigenvector with the highest eigenvalue. To create a new feature space, select the top k eigenvectors. The amount of variation (e.g., 95%) that is desired to be retained determines the value of k.

Reduces the number of features while keeping the most crucial information, which simplifies the dataset. lowers the processing expense of ensuing data processing jobs. Generates uncorrelated characteristics that have the potential to enhance the efficiency of specific machine learning techniques. Projects high-dimensional data into two or three dimensions to aid in its visualization.

3.5. Machine Learning Methods. Applying data mining techniques to educational data in order to predict academic performance and study student behavior is known as educational data mining, or EDM. Institutions can deliver individualized learning experiences, enhance educational achievements, and spot patterns by utilizing machine learning. The procedures for creating a machine learning model to forecast student performance are described in this framework.

3.5.1. Decision Trees. A decision tree is an arrangement that resembles a flowchart, with each internal node denoting a choice made in response to a feature (attribute), each branch denoting the decision's result, and each leaf node representing a class label (in this case, student performance). The routes from the root to the leaf show the guidelines for classification. Collect data about students, covering a range of aspects such as personal and academic history, attendance, behavior, and other pertinent characteristics.

Starting with the complete dataset and choose the feature (e.g., pass or fail) that divides the data into the most distinct classes. Information gain, entropy, and Gini impurity are frequently used criteria to determine the optimal split. The dataset was iteratively divided into subsets according to the chosen characteristic. The goal of each split is to produce subsets that are purer, which means that the subsets are supposed to only include data points from one class.

Student data from the past is used to train the decision tree. The framework discovers connections and patterns among the goal variable—such as grades or pass/fail status—and the characteristics of the inputs. By navigating the tree according to the student's characteristics, the decision tree system can forecast a new student's success.

4. Result Analysis. The efficacy evaluation of forecasting algorithms created with instructional data mining approaches is the main objective of the outcome analysis. Employing a variety of academic and demographic variables, these models seek to predict the performance of students. For a broad range of students, the collection contains educational records, records of attendance, information on demographics, and indicators of socioeconomic status. Academic results (tests, assignments, and final examinations), attendance records, involvement in extracurricular endeavors, educational levels, and other key factors are all examined.

The linear link between the observed and expected values is measured by the correlation coefficient, which also indicates its direction and intensity. There is a significant positive association when the value is 0.86. The average absolute difference between the expected and actual values is represented by the mean absolute error, or MAE. An average deviation of 18.92 units between the predicted and actual values indicates that the predictions are not accurate.

The square root of the average squared discrepancies between the expected and actual values is indicated by the Root-Mean-Squared Error (RMSE). The residuals' (prediction errors') standard deviation is displayed with a value of 24.31. A larger average error is implied by a higher value. The square root of the mean absolute variations between the expected and actual values is what is measured by the root-absolute error. The degree of prediction mistakes is represented by the value of 17.16; a smaller value indicates better results. Root Relative Squared Error (RMSE): Provides a normalized measure of error by reflecting the RMSE in relation to the range of observed values. The RMSE in relation to the data's magnitude is represented by the value of 19.51. In figure 4.1 shows the overall performance metrics of educational data prediction.

Out of the four classifiers, the Decision-Tree classifier has the highest accuracy, just over 95%. At about 85%, the K-NN classifier has the lowest accuracy.While it does not perform as well as the Decision-Tree or GA classifiers, the GA+K-NN classifier outperforms K-NN. The Decision-Tree classifier is the most accurate, followed by the GA, GA+K-NN, and K-NN classifiers, as this figure 4.2 graphically illustrates.

With the lowest RMSE, the GA+Decision-Tree model performs best with the least amount of error. The K-NN model performs the worst and has the most errors, as evidenced by its highest RMSE.The RMSE values of the Decision-Tree (DT) and GA+K-NN models are in the middle, with the Decision-Tree model outperforming GA+K-NN. The GA+Decision-Tree model is the most accurate, followed by the Decision-Tree, GA+K-NN, and K-NN regression models, as this chart 4.3 illustrates clearly.

5. Conclusion. The substantial potential of using data analytics to improve educational outcomes has been shown by the research on educational data mining for pupil achievement prediction. Through the implementation of diverse machine learning techniques and statistical methods on student data, this study has

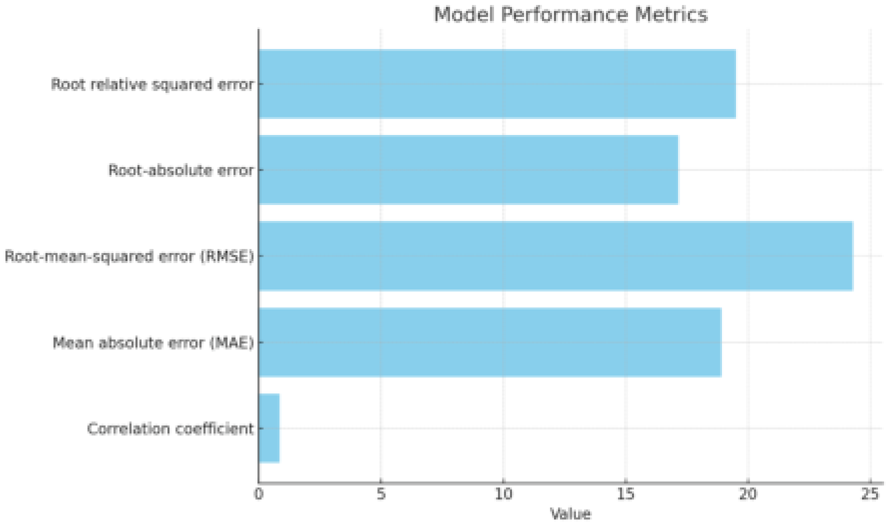


Fig. 4.1: Performance metrics

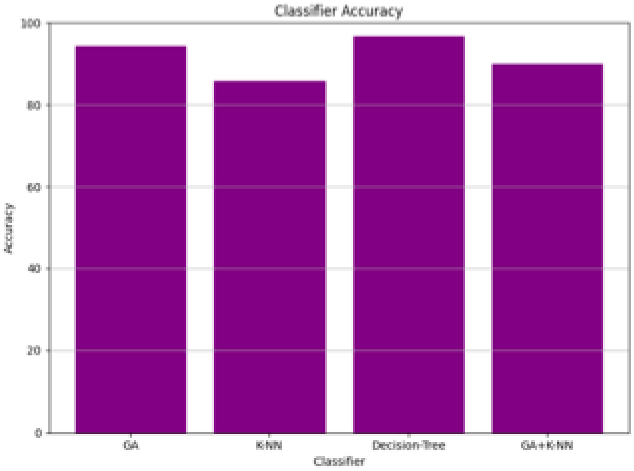


Fig. 4.2: Classification accuracy

effectively discovered pivotal aspects that impact academic achievement. With the help of the models for prediction created in this study, teachers will be able to proactively address any academic challenges that may arise and customize their lesson plans to meet the requirements of each unique student. The findings highlight the value of ongoing data gathering and analysis in learning environments. The knowledge gathered from these studies aids in the creation of individualized education programs in addition to providing insight into the behavior and learning behaviors of students. Additionally, the creation of curriculum, allocation of resources, and institutional decision-making can all be aided by the use of educational data mining, which will ultimately promote a more encouraging and productive learning environment. To further improve the models for prediction, future studies should concentrate on incorporating a wider range of data sources, such as behavioral and socioeconomic variables. To ensure that the advantages of data mining in education are realized without jeopardizing student privacy, ethical concerns about data security and privacy must also be taken into account. To sum up, data mining in education presents a viable way to use information-driven knowledge to raise the

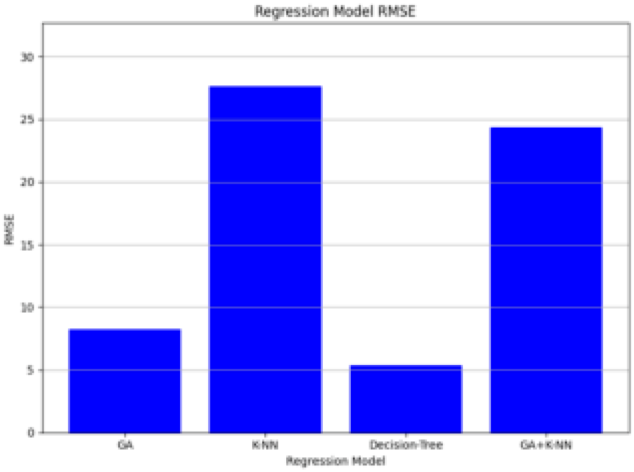


Fig. 4.3: Regression Model RMSE

achievement of students. Schools can raise academic expectations, improve learning experiences, and provide greater assistance for their students by utilizing analytics to predict outcomes.

Subsequent investigations can concentrate on investigating and incorporating increasingly sophisticated machine learning models, like deep learning architectures, ensemble techniques like Extreme Gradient Boosting (XGBoost) and Gradient Boosting Machines (GBM), and hybrid models that blend the advantages of many algorithms. These algorithms might perform better when managing intricate, large-scale educational statistics, resulting in more precise forecasts and profound understanding of student achievement.

REFERENCES

[1] F. A. Al-Azazi and M. Ghurab, *Ann-lstm: A deep learning model for early student performance prediction in mooc*, heliyon, 9 (2023).

[2] A. Alam, *Improving learning outcomes through predictive analytics: Enhancing teaching and learning with educational data mining*, in 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2023, pp. 249–257.

[3] ———, *The secret sauce of student success: Cracking the code by navigating the path to personalized learning with educational data mining*, in 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), IEEE, 2023, pp. 1–8.

[4] S. Albahli, *Efficient hyperparameter tuning for predicting student performance with bayesian optimization*, Multimedia Tools and Applications, 83 (2024), pp. 52711–52735.

[5] A. S. Alghamdi and A. Rahman, *Data mining approach to predict success of secondary school students: A saudi arabian case study*, Education Sciences, 13 (2023), p. 293.

[6] K. Aulakh, R. K. Roul, and M. Kaushal, *E-learning enhancement through educational data mining with covid-19 outbreak period in backdrop: A review*, International journal of educational development, 101 (2023), p. 102814.

[7] C. Baek and T. Doleck, *Educational data mining versus learning analytics: A review of publications from 2015 to 2019*, Interactive Learning Environments, 31 (2023), pp. 3828–3850.

[8] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, *Educational data mining to predict students' academic performance: A survey study*, Education and Information Technologies, 28 (2023), pp. 905–971.

[9] N. Bayes and B. A. Ningsi, *Performance comparison of data mining classification algorithms on student academic achievement prediction*, Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 6 (2023), pp. 29–39.

[10] Y. Chen and L. Zhai, *A comparative study on student performance prediction using machine learning*, Education and Information Technologies, 28 (2023), pp. 12039–12057.

[11] P. Guleria and M. Sood, *Explainable ai and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling*, Education and Information Technologies, 28 (2023), pp. 1081–1116.

[12] C. Huang, J. Zhou, J. Chen, J. Yang, K. Clawson, and Y. Peng, *A feature weighted support vector machine and artificial neural network algorithm for academic course performance prediction*, Neural Computing and Applications, 35 (2023), pp. 11517–11529.

[13] S. Hussain and M. Q. Khan, *Student-performulator: Predicting students' academic performance at secondary and inter-*

mediate level using machine learning, Annals of data science, 10 (2023), pp. 637–655.

[14] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, *Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms*, Education and Information Technologies, 28 (2023), pp. 9655–9684.

[15] S. Mallak, M. Kanan, N. Al-Ramahi, A. Qedan, H. Khalilia, A. Khassati, R. Wannan, M. Mara'beh, S. Alsadi, and A. AlSartawi, *Using markov chains and data mining techniques to predict students' academic performance*, (2023).

[16] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, *Classification and prediction of student performance data using various machine learning algorithms*, Materials today: proceedings, 80 (2023), pp. 3782–3785.

[17] M. H. B. Roslan and C. J. Chen, *Predicting students' performance in english and mathematics using data mining techniques*, Education and Information Technologies, 28 (2023), pp. 1427–1453.

[18] X. Wang, Y. Zhao, C. Li, and P. Ren, *Probsap: A comprehensive and high-performance system for student academic performance prediction*, Pattern Recognition, 137 (2023), p. 109309.

[19] T. Wongvorachan, S. He, and O. Bulut, *A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining*, Information, 14 (2023), p. 54.

[20] O. R. Yürüm, T. Taşkaya-Temizel, and S. Yildirim, *The use of video clickstream data to predict university students' test performance: A comprehensive educational data mining approach*, Education and Information Technologies, 28 (2023), pp. 5209–5240.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments
Received: Jul 23, 2024
Accepted: Aug 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 $\LaTeX2_\varepsilon$ 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.