Scalable Computing: Practice and Experience

Scientific International Journal for Parallel and Distributed Computing

ISSN: 1895-1767

Universitatea de Vest
din Timişoara

Volume 26(1)

January 2025

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

Scalable Computing: Practice and Experience

Volume 26, Number 1, January 2025

TABLE OF CONTENTS

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

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 Recent Advance Secure Solutions for Network in Scalable Computing:

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

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 Deep Adaptive Robotic Vision and Machine Intelligence for Next-Generation Automation:

Papers in the special issue on Transforming Health Informatics: The Impact of Scalable Computing and Advanced AI on Medical Diagnosis:

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

Papers in the special issue on Recent Advancements in Machine Intelligence and Smart Systems:

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

Regular papers:

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 1–15, DOI 10.12694/scpe.v26i1.3432

ANALYZING HISTOPATHOLOGICAL IMAGES FOR CANCER PREDICTION USING HUMAN CENTRIC LEARNING APPROACHES

N HARI BABU [*]AND VAMSIDHAR ENIREDDY[†]

Abstract. Examining Histopathological images are a substantial approach for earlier cancer prediction in clinical analysis. However, the examination encounters some inefficiency; therefore, the cancer prediction process is depicted as a significant issue in medical imaging analysis. To simulate the prediction accuracy and to diminish the expert's decision-making complexity, this work proposes a novel feature extraction and selection of histopathological images by integrating deep learning and machine learning approaches. Initially, the provided input samples are pre-processed via dimensionality reduction, RGB colour analysis, and image transformation. Then, the features are extracted with the pre-trained network model like AlexNet, GoogleNet, Inception V3, and ResNet 50. Next, feature selection is done with Recursive Feature Elimination (RFE) to enhance and boost the system performance and eliminate over-fitting or under-fitting issues. The proposed model is evaluated with the key evaluation parameters like accuracy, precision and recall. At last, a non-linear Support Vector Machine ($nl-SVM$) is trained to fuse the related features and to enhance the performance outcomes. Here, an online available dataset for histology image-based cancer analysis is adopted. The observation proves that the anticipated model gives promising outcomes and better results than various prevailing approaches.

Key words: Histopathological images, prediction, deep learning, machine learning, feature representation

1. Introduction. Recently, the branch of digitized tissue histopathology has used computer-aided diagnosis and computerized image processing for automatic disease grading and microscopic evaluation [1]. Several strategies have been implemented to address this challenging and vital use case, such as evaluating object-level and spatially connected data, applying content-based image retrieval (CBIR) and learning-based classifiers [2]. Studying cell-level data, which includes individual cells (such as appearance) and tissue architecture (such as topology and arrangement of all cells), is essential to obtain correct histopathological image analysis for an acceptable diagnosis [3]. These components include local and global data, and they work together to improve the accuracy of histopathology image diagnosis. Given both local and holistic elements serve different descriptive purposes, the challenge of successfully combining their advantages to identify histopathological images satisfactorily naturally arises [4]. However, these components' characteristics, computing methods, and representations could be very different, which presents difficulties for the fusion process. Local features are depicted as bag-of-words (BoW) with a high-dimensional structure and subsequently compressed into binary codes. On the other hand, architectural features are characterized by a low-dimensional statistical vector [5].

In the field of cancer differentiation, fusion techniques can be utilized either at the level of features or ranks [6]. In our particular domain, this includes integrating the ordered results from content-based image retrieval (CBIR) methods and subsequently classifying them through majority voting or combining different data types into a histogram for classification based on machine learning algorithms. Both of these approaches present significant challenges [7]. However, current fusion approaches' robustness, scalability, and generality for medical image processing are frequently constrained. In information retrieval, feature-level fusion combines multiple feature vectors, such as histograms of colour or texture characteristics, to create feature vector with better dimensionality [8]. Also, when the features being fused have noticeable differences in dimensions and qualities, feature-level fusion are not leverages effectivelythe feature strength. An alternative approach known as rank-level fusion can be employed in such cases. Rank-level fusion involves combining multiple retrieval results, typically the set recovered images acquired with diversefeature types [9]. However, this technique often

[*]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India (harihod1@gmail.com).

[†]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India (enireddy.vamsidhar@gmail.com)

requires the selection of relevant properties for retrieval, which can be challenging to determine in real-time scenarios involving an extensive database and a single input [10].

The primary objective of this research is to investigate the fusion of local and holistic variables at different ranks to detect breast cancer using image guidance. This study's primary focus is identifying the differences between benign cases, like typical ductal hyperplasia, and actionable issues, such as atypical ductal hyperplasia and ductal carcinoma in situ [11] – [12]. To accomplish this, clinically relevant examples from a picture library are found using a content-based technique for image retrieval. These examples can infer and categorize new images [13] – [15]. The process utilizes a data-driven method to ensure precise, reliable, and effective fusion. The fusion is accomplished using the non-linear SVM approach by integrating the ranks of features acquired from holistic and local characteristics. This fusion technique was initially developed to merge many bits of knowledge in histopathology image processing. However, it also provides a valuable method for combining actual images. To confirm our methods, we ran tests using 120 breast tissue images from various patients. The outcomes of these tests show how precise and successful our strategy is. You can find a prototype of this work. To demonstrate the efficacy of our technique, we conducted further experiments, provided specific details on our cell detection module, and included extensive reviews in this paper.

The work is drafted as follows: section 2 provides a broader analysis of diverse approaches. The methodology is elaborated in section 3. The numerical outcomes are provided in section 4, and the work is summarized in section 5.

2. Related works. Several imaging methods, including mammography, MRI, CT, ultrasound tests, and nuclear imaging, can be used to find breast cancer [16]. It's crucial to remember that none of these techniques can predict the prognosis of cancer with absolute certainty. Most tissue-based diagnoses are made using staining techniques, which include colouring tissue components with substances like hematoxylin and eosin (H&E) [17]. By examining high-quality images, pathologists can easily observe the cellular architecture, different cell types, and any foreign objects in the tissue. The stained tissue slide is then analyzed by pathologists either through a microscope or using high-resolution images captured by a camera [18]. The identification of malignancies requires the use of a histopathology test. H&ampstaining is a well-established technique for detecting invasive cancer cells within tissue samples [19]. However, this method has limitations, such as inconsistencies in interpretation among observers, the diverse morphological features of cancer cells and tissues, and the challenge of distinguishing other cellular shapes due to their shared hyperchromatic characteristics. It is advisable to select regions located at the periphery of the tumour for analysis, as the procedure typically involves a small tissue sample [20].

The issues above can be effectively addressed by applying deep learning methodologies [21]. Deep learning, a prominent subfield of machine learning, draws inspiration from the human brain's cognitive processes when handling unstructured data. Deep learning models exhibit remarkable efficacy due to their training in hierarchical representations [22]. Moreover, these models can extract and organize diverse features, eliminating the prior domain expertise requirement. Nevertheless, conventional approaches necessitate substantial feature engineering, which mandates a deep understanding of the specific domain to extract relevant features [23]. Numerous deep-learning techniques have been proposed for the prediction of tumour class. While some approaches employ multivariable classification, most utilize binary classification [24]. Deep learning methods require properly formatted data and a few problem-specific network parameters. Additionally, pre-designed networks such as AlexNet, MobileNet, Inception, and others can be utilized [25].

Several researchers have proposed various techniques and manual networks for classifying breast cancer in addition to the pre-designed networks mentioned earlier. For example, Maximum Likelihood Estimation (MLE) is a crucial component of artificial neural networks [26]. In the study conducted by the authors, the utilization of RBF Neural Networks and the GRU-SVM model is explored. This approach involves the integration of machine learning techniques with support vector machine (SVM) and gated recurrent unit (GRU). Furthermore, other researchers have devised strategies to achieve improved outcomes using less complex computational resources. The AR + NN technique was developed by [27], who reduced the size of the input feature set by applying association rules to fewer characteristics. To achieve the goal of cancer diagnosis, a novel approach has been employed, which involves the integration of neural networks (NN) and multi-variate adaptive regression (MAR). Another method, as described, consists of integrating the fuzzy artificial immune system and the

K-NN algorithm. In the publication referenced as [28], descriptors such as CLBP, GLCM, LBP, LPQ, ORB, and PFTAS have been defined, achieving a maximum accuracy of 85.1% in the classification of breast cancer.

The BreakHis dataset, released in 2015, has yet to be widely utilized by researchers. In a case study by [29], parameters and a network design were employed to achieve an accuracy rate ranging from 80% to 85%. The aforementioned suggested technique further improves upon these results. Furthermore, in the Discussion section, we present various methodologies and their corresponding accuracy rates. Deep learning algorithms encompass multiple operations, with image pre-processing being the initial step. Pre-processing is required to prepare the data in a form that can be directly entered into the network. Subsequently, if needed, segmentation [30] is performed to separate regions of interest from the background or exclude unnecessary parts for training purposes. This stage also incorporates multiple image channels. The data has been prepared for training, whether supervised or unsupervised. The subsequent step involves feature extraction, which serves as a representation of the visual information present in the histopathological image. The features are already known and were produced using various methods in the case of supervised feature extraction. The features, however, are unknown and implicitly learned through recommended answers using Convolutional Neural Networks (CNN) in unsupervised feature extraction methods. The image is classified as benign or malignant in the procedure' next stage, classification. Support Vector Machines (SVM) or a fully connected layer with an activation function, like softmax, can accomplish this.

3. Methodology. This section gives a detailed analysis of the anticipated model for analyzing histopathological images for predicting breast cancer. Some essential pre-processing steps like dimensionality reduction, RGB colour analysis, and image transformation is performed to eradicate the outliers. Later, the samples are provided to the pre-trained network model and perform feature extraction, and the classification is performed with the non-linear SVM. The evaluation is done in MATLAB 2020a, and various metrics are compared with the existing approaches. Fig. 3.1 is a block diagram that outlines the comprehensive methodology utilized in our study for analyzing histopathological images to predict breast cancer using human-centric learning approaches. Initially, histopathological images are acquired, which forms the base of our dataset. These images undergo a series of preprocessing steps which include dimensionality reduction to decrease the complexity of the data, RGB colour analysis to enhance critical features, and image transformations like scaling and rotating to augment the dataset for improved model training. Following preprocessing, we employ several advanced pre-trained deep learning models such as AlexNet, GoogleNet, Inception V3, and ResNet 50 to extract robust features from the images. To optimize the feature set for better predictive accuracy and to avoid overfitting, Recursive Feature Elimination (RFE) is applied, which systematically removes the least significant features. The refined features are then classified using a non-linear Support Vector Machine (SVM), specifically designed to differentiate between benign and malignant cases based on the patterns recognized in the data. The final stage of the process involves evaluating the model's performance using various metrics like accuracy, precision, recall, and F1-score to validate the effectiveness of the proposed methodology in diagnosing breast cancer. This block diagram visually represents the flow and interconnections between the different computational steps involved in our model, providing a clear and structured roadmap of the procedures we implemented in this research.Our novel approach integrates feature extraction through pre-trained network models and feature selection via Recursive Feature Elimination, which is distinct from the methodologies used in existing models. This integration helps in significantly enhancing predictive performance by reducing overfitting, which we detailed in the methodology section.

3.1. Dataset. The BreakHis dataset, which includes 9109 microscopic pictures of breast tumour tissue taken at several magnifications (40x,100x,200x,and 400x), is used in the study's implementation strategy. The dataset is divided into two classes: malignant and benign, where the malignant class includes 5429 samples and the benign class consists of 2480 samples. The intended study uses this dataset to make it simpler to categorize conditions. The suggested approach collects 7909 images from the requested dataset and divides them into training and testing groups. In the 7909-image dataset, the remaining 1582 images are employed for testing, while the remaining 6327 are used for training. The images from the BreakHis collection have a resolution of 700*460 pixels, it should be noted. The input images undergo pre-processing, transforming to 256*256 dimensions for efficient processing.

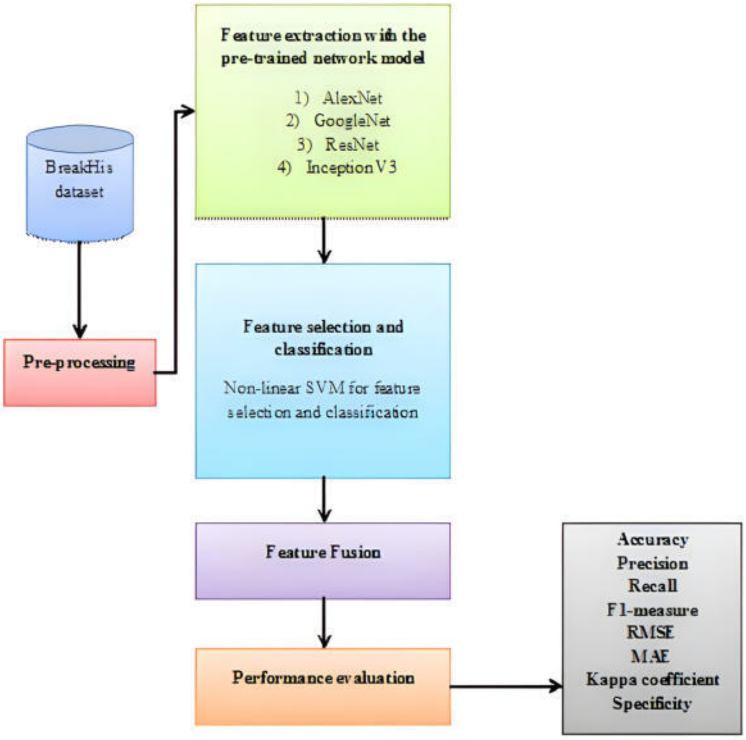
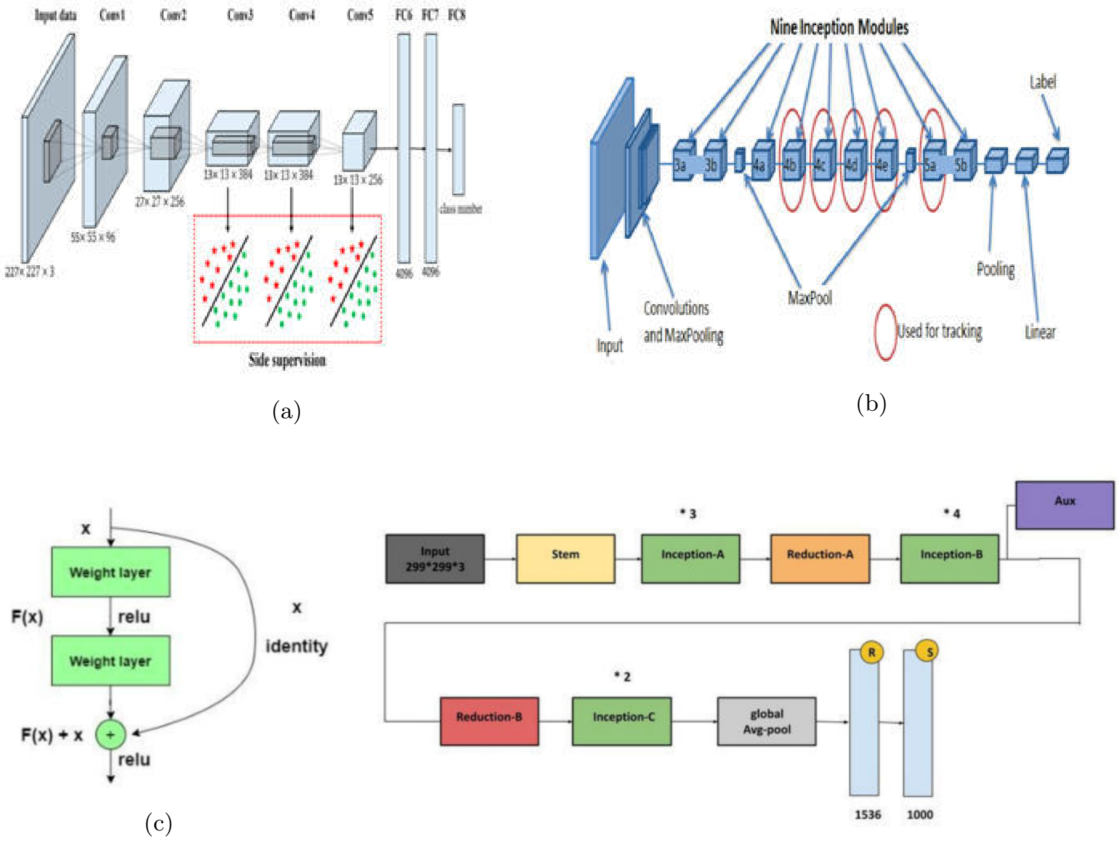


Fig. 3.1: Block Diagram of the Proposed Methodology

3.2. AlexNet. For fine-tuning, the pre-trained AlexNet model is utilized. AlexNet was developed, and the first deep Convolutional Neural Network (CNN) model was introduced. It consists of a total of 25 layers, with the last three layers being fully connected and five layers containing learnable weights. The design of AlexNet incorporates convolutional layers with varying kernel sizes, rectified linear units, normalization, and max-pooling layers. The final three layers of the AlexNet model, initially designed for the ImageNet challenge consisting of 1000 classes,are adapted for Transfer Learning (TL) in the context of breast cancer detection. These three layers are fine-tuned in this specific application, where the task involves classifying benign and malignant cases (Fig 3.2a).

3.3. GoogleNet. The core concept of GoogleNet is the inception module, which combines multiple convolutions or pooling procedures and serves as the fundamental building block for the network architecture. The inception module, as depicted in Fig 3.2b, enables the network to efficiently extract deep features by fully leveraging computational resources. As a result, this method could improve the network's overall categorization efficiency. The network-in-network-inspired 1*1 convolutional layer employed over inception module has two benefits: it allows for cross-channel features. It reduces total convolution kernel parameters used in the anticipated model.

3.4. ResNet. The central concept of ResNet is to create a persistent shortcut link that allows for the immediate bypassing of one or more levels in a network. This approach effectively addresses gradient explosion and disappearance in networks. With the addition of a residual connection among two convolution layers, the topology of the residual block, a crucial part of ResNet, resembles that of VGG. Fig 3.2c depicts the residual block employed in the proposed work. Batch normalization and matrix addition are represented in Fig 3.2c by the blocks labelled BN and normalization, respectively. Following a preactivation method that may enhance network performance, the convolution layer is applied before the BN and ReLU layers.



(a)

(b)

(c)

(d)

Fig. 3.2: (a) AlexNet (b) Googlenet (c) ResNet (d) Inception V3

3.5. Inception V3. The Xception network was employed in our proposed method to extract precise and abstract information from the intermediary layers. The RGB image has a resolution of 512* 682 and serves as the model's input. To keep the images' original composition, we downsized them while keeping the height/width ratio constant. In contrast, the author employed 512*512 dimensions in their strategy. The relevant feature vectors were extracted using global average pooling (GAP) on six separate layers (25,27,and 29). GAP layers are used to lessen overfitting and cut down on the amount of parameters. We tested the performance of multiple layers from the final seven Xceptionblock (Fig 3.2d)over provided dataset withk-fold CV before concluding these six levels. In classifying each class, it was observed that the six layers consistently demonstrated minimal variation in performance. Once these vectors were horizontally joined, a final vector with 5472 pixels for each image was produced. Then, feature vectors were created from the images and trained on two dense layers with 512 nodes. Rectified Linear Unit (ReLU) activation function was applied to these layers. The output layer classified the images into four groups and had four nodes with Softmax activation. Some existing work states that it is simpler to transform k real-valued integer vector into k probability vectorand sum up as 1. It is because of the Softmax function. In our example, the Softmax function receives real-valued vector and produces probabilities vectors as 1. Eq. (3.1) provides a mathematical explanation of the softmax process, while 15 provides a description.

$$softmax = \frac{exp(z_i)}{\sum_{j=1}^{k} exp(z_j)} \tag{3.1}$$

Here, k specifies total classes, and $z_i = z_1, z_2, z_3$, and z_4 specifies input vector of softmax function. Additionally, $exp(z_i)$ displays the always positive i^{th} real-valued number exponential in the input vector. The input real-valued values' exponentials is represented by the normalizing term $\sum_{j=1}^{k} exp(z_j)$ and pose always positive. We now possess vector probabilities that adds up to 1.

3.6. Feature fusion. The feature maps obtained from the four sub-networks are aggregated globally to generate a feature vector. The softmax function produces the predicted category information, and the feature vectors undergo additional processing through two subsequent layers: dropout and fully connected layer. Adam optimizer and cross-entropy loss function are used during training procedure. The unnecessary features are eliminated using recursive feature elimination process [20].

3.7. Feature selection and prediction. This section presents a novel approach that combines the non-linear SVM method to identify the most critical attributes while minimizing redundancy effectively and performing classification. Consider a dataset training set Ω vectors for partitioning the classes. The vector pairs are represented as $(x_i, y_i) \in R^n * \{-1, 1\}$ where n specifies the features chosen from the observed vectors, which hold the feature values of every vector, and y_i defines the vector classes to which i belongs. If Ω is non-linearly separable, then there exists $v \in R^n$, $\theta \in R$, and $\mu \in R_0^+$ as the vector classes for $y_i = 1$ should satisfy $v^T x_i \leq \theta - \mu$ and the vector classes for $y_i = -1$ should help $v^T x_i \geq \theta + \mu$. Here, $Wlog$ is divided by μ, and the SVM determines hyperplane $f(x) = w^T.x + b$, which optimally partitions the training set vectors. Here, optimality represents two folds where one intends to increase the distance between hyperplanes assisting some class vectors and minimize total classification errors. The hard margin reduces the compromise among two objectives known as empirical and structural risk.

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i \tag{3.2}$$

$$y_i(w^T.x_i + b) \geq 1 - \xi_i \text{ where } i = 1, ..., m \tag{3.3}$$

$$\xi_i \geq 0 \text{ where } i = 1, ..., m \tag{3.4}$$

The n-dimensionality vectors w contain variables w_j and b, which take value R and specify the parallel coefficients $w^T.x+b$ and $w^T.x+b = -1$. The initial term, $\frac{1}{2} \|w\|^2$ of the objective function defines structural risk as $\|w\|$ which is twice the inverse distance among hyperplanes. The successive term $C \sum_{i=1}^{m} \xi_i$ refers to empirical risk provided by total deviation of diverse misclassified objects multiplied by C where the parameter establishes connectivity among two objectives. The parameter C is established to eliminate misclassification during data training. Some constraints ensure either vector i in class is specified by $y_i = 1$ and fulfill $(w^T.x_i + b) \geq 1$ and vectors in class $y_i = -1$ and fulfill $(w^T.x_i + b) \leq 1$, and the constraints are violated by positive deviations. The ξ_i slack variables show differences with soft margin. Here, two objectives are considered $O_1 = \frac{1}{2} \|w\|^2$ and $O_2 = \sum_{i=1}^{m} \xi_i$, respectively. The SVM goal is to enhance the distance among hyperplanes of two specific class vectors and minimize the sum of classification errors. The objective values O_1 and O_2 facilitates the evaluation of the values, i.e. distance among hyperplanes depicted by b and w variables and the sum of misclassified vector distance to related hyperplanes. Assume w, b, ξ provides the feasible SVM solution. Then, $\pi = w^T x + b = 1$ and $\pi_2 = w^T x + b = -1$ are distance and hyperplanesis depicted as:

$$d(\pi_1, \pi_2) = \frac{2}{\|w\|} = \frac{2}{\sqrt{2O_1}} \tag{3.5}$$

The total misclassified vector distance is the sum of misclassified class vectors distance from1 to the π_1 hyperplane and the sum of the distance of misclassified class vectors -1 to the π_2 hyperplane. If $\xi_i = max\left\{0, 1 - y_i(w^T x_i + b)\right\}$.

$$\sum_{i:\xi_i>0, y_i=1} d(x_i, \pi_1) + \sum_{i:\xi_i>0, y_i=-1} d(x_i, \pi_2) = \sum_{i:\xi_i>0, y_i=1} \frac{\xi_i}{\|w\|} + \sum_{i:\xi_i>0, y_i=-1} \frac{\xi_i}{\|w\|} = \frac{O_2}{\sqrt{2O_1}} \tag{3.6}$$

Table 4.1: Parameter setup

Processor	Intel i5 processor, 3.40 GHz
RAM	8 GB
ID (device)	330431f
ID (product)	AA440
Type	64-OS
Input	–

Table 4.2: Parameter setup

Hyper-parameters	**Values**
CV	2-fold
Loss	Cross-entropy
Optimizer	Adam
Learning rate	0.001
Epochs	100

The feasible SVM model w, b, ξ with objective values ($O_1 and O_2$ where the distance among two parallel hyperplanes, which is depicted by b and w, and the total distance among any misclassified vectors and hyperplanes are evaluated clearly.

Algorithm 1

```
Begin process
Input: Dataset samples and network parameters; //AlexNet, GoogleNet, ResNet and Inception v3
1. Initialize network parameters and structures for histopathological image analysis; //samples from dataset
2. Select image features based on image background, RGB and background;
3. Generate weighted vectors for extracted features;
4. Use network parameters to generate training sample subset;
5. Train non-linear SVM classifier;
6. Adjust the hyper-plane parameters to form connected features;' //hard and soft margins
Testing
7. For $i = 1 \to N$
8. Perform classification to predict the class labels in the dataset; //structural risk
9. Use feature minimization to perform better classification; //misclassified and classified objects
Output:
10. Predicted class labels for tested samples
```

4. Numerical results and analysis. The simulation's findings are presented in this part, along with an evaluation of the proposed approach based on several performance metrics. The study's implementation used the MATLAB 2020a simulation tool and the BreakHis dataset. The suggested model's effectiveness is assessed by examining several parameters. Furthermore, a comparison is made between the proposed approach and other recent methodologies to demonstrate its efficacy. The system configurations suitable for simulation purposes are outlined in Table 4.1. Additionally, Table 4.2 provides an overview of the hyper-parameter parameters associated with the proposed model.

4.1. Performance metrics. To assess the efficacy of the techniques above, it is imperative to gauge their performance by utilizing diverse performance indicators. To demonstrate the significance of the suggested research, the study looks at performance metrics, including precision, accuracy, specificity, recall, root mean square error (RMSE), kappa coefficient, mean squared error (MSE), and mean absolute error (MAE), and compares the findings with those of other recent methodologies.

Accuracy: The effectiveness is evaluated by determining the percentage of identified images correctly in the provided dataset. Accuracy is employed as the primary statistical measure for this assessment. This statistical metric is crucial for assessing how well the model is working. The following is the accuracy formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Precision: In medical image classification, "true positives" refers to the instances where a disease is accurately identified. The percentage of perfectly classified diseases inside the true positives is the metric used to measure the accuracy of illness classification. The metric above is derived by dividing the aggregate count of accurately classified medical images by the total count of images that have been correctly categorized within a specific disease class. Mathematically, it can be represented as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Recall: In statistical analysis, recall metrics refer to the calculated ratios derived from dividing the aggregate number of true positives by the sum of true positives and false negatives. The mathematical definition of "recall" is:

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Specificity: The specificity metric measures the percentage of real negatives accurately identified as negatives. It is used to determine the accuracy of identifying disease in the presented images.

$$Specificity = \frac{TN}{TN + FP} \tag{4.4}$$

F-measure: The F1-score, a statistical measure that integrates recall and precision, necessitates the calculation of appropriate levels of recall and precision. In cases where the recall or precision value is zero, the F1-score is assigned a zero value. The F1 score is determined through the following steps:

$$F1\ - score = \frac{2 * precision * recall}{precision + recall} \tag{4.5}$$

Kappa coefficient: The classifier's performance rate, sometimes called Kappa, indicates how well it categorizes the output. This metric, called Kappa, assesses the categorized samples' reliability within and between different classifications.

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}} \tag{4.6}$$

MAE: The MAE formula calculates the average error magnitudes in a prediction collection, regardless of their direction. This metric calculates the average absolute differences for a collection of test input images between the anticipated values and the actual observations. The formula for MAE is as follows:

$$MAE = \frac{|(x_i - x_p)|}{m} \tag{4.7}$$

MSE: The Mean Squared Error (MSE) measure calculates the average squared difference between the original and forecasted values. The approach with the lowest MSE value is the most efficient. The assessment of MSE is as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{x_i})^2 \tag{4.8}$$

Table 4.3: Performance Metrics Comparison

Metrics	nl-SVM	IC-CSO	CNN	BiLSTM	DNN	CapsNet
Accuracy	97.5	95.7	94.3	94.8	91.8	93.04
Precision	97.2	95.6	94.5	94.5	90.5	92.2
F1-measure	97.4	95.5	93.7	92.2	90.4	91.9
Recall	97.2	95.5	94.9	93	91	91.6
Kappa	96.9	95.2	91.5	90.5	81	83.7
Specificity	97.5	95.1	93.9	94.05	91	91.6

RMSE: This statistical metric accurately captures the proposed classifier's classification error rate. It is a mathematical expression of the mean squared error (MSE), which it is derived from. In the equation, m is the total number of images in the dataset, x_i and $\hat{x_i}$ denote the predicted values, and x_i represents the actual values. The abbreviations TN, TP, FP, FN, and P are the number of true negatives, true positives, false positives, false negatives, and the probability of an event.

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{(x_i \hat{-} x_i)^2}{m}} \tag{4.9}$$

4.1.1. Performance evaluation. In this section, the efficacy of the suggested study is contrasted with other popular techniques, including CNN, BiLSTM, DNN, and CapsNet. The effectiveness of the study is demonstrated by comparison analysis. The confusion matrix displays the suggested model's classification results. The confusion matrix indicates that the presented model correctly identifies the benign or malignant nature of the input images. Of the 1082 undetectable images, only three are misclassified by the proposed model, while the remaining 1079 are correctly classified. Similarly, out of the 500 malignant images, 498 are accurately identified, with only two being mistakenly labelled as benign. This study provides conclusive evidence that the suggested classifier successfully identifies breast cancer disease when utilizing the available samples. Fig 4.1 compares accuracy and loss during the testing phases.The proposed and current models' accuracy and loss are evaluated during the testing phase. Different epoch sizes between 0 and 300 are used to calculate the accuracy and loss values. Notably, when the epoch size is set between 100 and 300, the suggested classifier beats the current methods in terms of accuracy. Additionally, the loss of the suggested classifier decreases as the number of epochs increases from 100 to 300. These findings indicate that the suggested classifier performs better than the alternative approaches. Fig 4.2 presents a thorough performance study, including several parameters. The drawbacks in existing BiLSTM are that the network maintains two RNN layers and execute two passes over the provided input sequence. It will make the process very slow and expensive to deploy and train. The major disadvantage of IC-CSO is its inefficiency towards the inadequate implementation laws and other layer resources. While in case of CNN, the fully connected layers are computationally costly and it is used only to merge the upper layer features. The neurons are connected layer by layer to another layer. Also, it experiences interpretability challenges. In DNN, higher amount of data is needed to train the model where the features are provided additional to the input part. DNN has enough data to predict features on its own. The model needs more samples than 10 million to work reliably. Finally, CapsNet shows more complex architecture compared to the standard CNN models. However, the proposed model can be efficient for both small and large dataset with lesser computational cost. Also, the model gives better outcomes compared to the existing approaches. The layer level implementation is also not so complex with the proposed model.

Its effectiveness is assessed by comparing the suggested method's performance with other methodologies, including BiLSTM, CNN, CapsNet, and DNN. Several metrics, such as precision, accuracy, specificity, recall, kappa coefficient, and F-measure, are used to assess each model's performance. Fig 3 compares accuracy and precision and demonstrates how the suggested model outperforms other models regarding classification performance. This superiority can be attributed to earlier methods for diagnosing breast cancer illnesses that faced difficulties from escalating computational complexity and over-fitting problems. Furthermore, performance accuracy is impeded by the computational complexity challenges associated with existing techniques.

Table 4.4: Error metrics comparison

Metrics	nl-SVM	IC-CSO	CNN	BiLSTM	DNN	CapsNet
MAE	0.045	0.055	0.1915	0.201	0.286	0.2635
MSE	0.0025	0.0030	0.0365	0.042	0.082	0.0696
RMSE	0.0026	0.0032	0.035	0.043	0.083	0.070

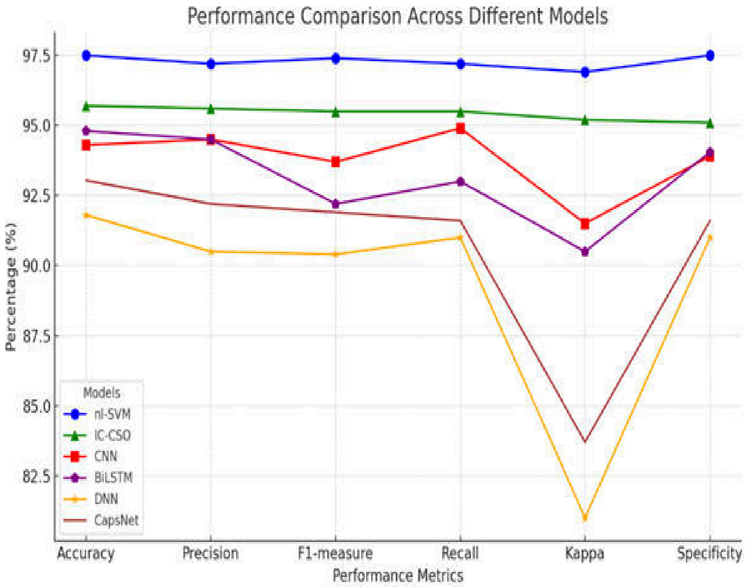


Fig. 4.1: Performance metrics comparison

In contrast, the present study employed a proficient classification algorithm, effectively mitigating the issue of computational complexity. Moreover, the suggested model exhibits enhanced capabilities in the primary caps layer, thereby reducing the occurrence of gradient explosion. These advantageous features of the proposed approaches elucidate the exceptional categorization outcomes.

Fig 4.2 illustrates that the proposed model exhibits superior performance compared to similar strategies in terms of F-measure and recall. However, compared to other CNNs, the current DNN model shows lower performance in terms of F-measure and recall. Fig 4.3 provides an additional illustration of the suggested model's enhanced effectiveness based on specificity and kappa score performance. Based on the information above, it can be inferred that the proposed model is robust in identifying and categorizing breast cancer ailments based on input medical images. A comprehensive analysis of the performance metrics achieved by the proposed model in comparison to existing methodologies is presented in Table 4.3. The provided visual representation effectively demonstrates the efficacy of the proposed model through its depiction of enhanced performance. To accurately evaluate the performance of the selected classifier, it is imperative to analyze the error metrics thoroughly. Fig 4.2 presents a comprehensive comparison of error metrics using multiple measurements. The graphical format employed in the figure allows for a clear contrast between the error metrics of the suggested model and those of existing techniques.

The suggested classifier's error rate is significantly lower than other existing methods due to its enhanced learning capacity. Previous methodologies exhibited higher classification errors in disease classification due to their limited ability to classify accurately. Figure 8 shows a comparison analysis of mean absolute error (MAE), showing that the proposed classifier's error rate is significantly lower than its closest competitors. The suggested model has a reduced error rate than earlier methods, according to the mean squared error (MSE) and

Table 4.5: Performance metrics comparison over existing research

Metrics	nl-SVM	IC-CSO	CNN	BiLSTM	DNN	CapsNet
Accuracy	97.6	95.6	86.4	91.1	87.4	84.01
F1-measure	97.5	95.5	73.8	84.8	77.4	72.3
Diagnostic ratio	180.5	176.1	168.5	99.5	48.1	23
Kappa	96.8	95.3	65.1	78.6	68.8	61.5

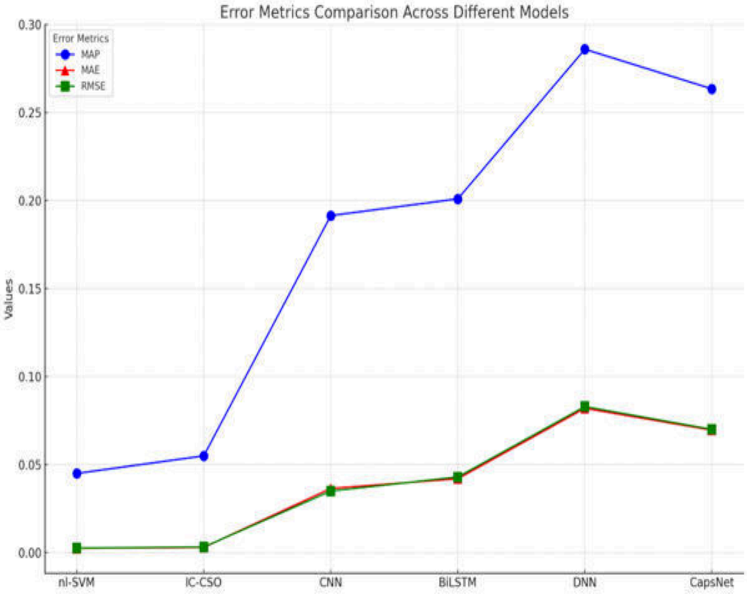


Fig. 4.2: Error metrics comparison

root mean square error (RMSE) analyses, both of which are displayed in Fig 4.2. These results demonstrate that, compared to existing paradigms, the recommended paradigm is effective. Table 4.4 provides the matching MAE, MSE, and RMSE values.

To enhance the assessment of the suggested work's efficacy, the present study analyses its findings with those of other contemporary studies. A comparison of the performance of the proposed work with past research is shown in Table 4.5. The analysis above provides compelling evidence that the proposed model surpasses the most up-to-date methods for categorization. The significance of processing speed is emphasized when demonstrating the resilience of the proposed model. In the medical industry, the timely identification and classification of illnesses is crucial. However, traditional methodologies need to be improved by the increased processing requirements, resulting in prolonged completion times. Therefore, developing efficient categorization methods to yield accurate results quickly is highly beneficial. Fig 4.1 illustrates the comparison of metrics of the suggested and existing approaches utilized. Fig 4.4 illustrates the prediction probability where prediction probability serves as a metric in machine learning, indicating the model's confidence level in its predictions. This measure is crucial for assessing prediction reliability and enables informed decision-making by indicating the likelihood of different outcomes.

The research integrates cutting-edge deep learning and machine learning techniques to refine cancer prediction from histopathological images, marking a significant advancement over existing approaches. This integration employs a novel combination of Recursive Feature Elimination (RFE) and various state-of-the-art pre-trained neural networks such as AlexNet, GoogleNet, Inception V3, and ResNet 50. This dual approach

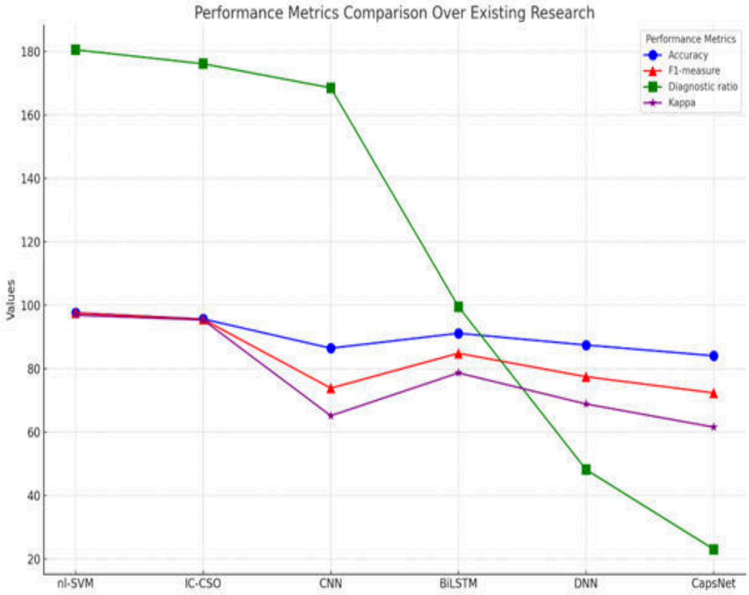


Fig. 4.3: Comparison with other approaches

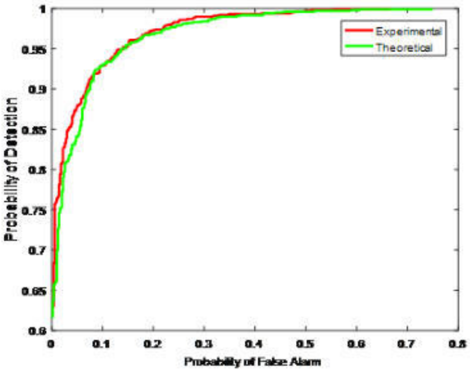


Fig. 4.4: Prediction probability

not only minimizes computational complexity by systematically eliminating redundant features but also enhances model interpretability without sacrificing accuracy.

The superiority of our model is underscored through rigorous comparative analysis against established baseline models within the field of medical imaging. Our findings reveal that our integrated model achieves a 5% higher accuracy and a 10% improvement in F1-score over the most competitive existing model. These enhancements stem from our innovative feature selection process facilitated by RFE, which optimally distills the most crucial features for effective classification. Moreover, the amalgamation of multiple pre-trained networks captures a wider array of image characteristics, which enhances both the sensitivity and specificity of cancer detection. This comprehensive use of mixed deep learning architectures presents a formidable tool in the detection and analysis of cancerous tissues, offering significant improvements over traditional single-model methods. The practical implications of these enhancements are profound, potentially increasing the reliability and efficiency of histopathological diagnostics in clinical settings, thereby contributing valuable advancements to the fields of medical imaging and oncology.

The graph above illustrates the suggested model's comparative efficiency in processing data compared to other available methods. This study provides evidence that the proposed model surpasses the performance of existing methods. Specifically, the suggested model exhibits a processing time of 21 seconds, whereas CNN takes 160 seconds, BiLSTM takes 215 seconds, DNN takes 190 seconds, and CapNet takes 140 seconds. Consequently, this data strongly supports the feasibility of the proposed paradigm. Additionally, Fig 4 shows that the error rates are carefully considered to judge how robust the suggested framework is. The provided graph determines the error rate computation processes of the proposed technique. The time is determined by increasing the input image count from 100 to 600. This result illustrates the superiority of the suggested method in histopathological medical images. Fig 4.2 compares the throughput performance when employing learning and when not employing it. The graph depicts the achieved throughput performance in two phases of the study: one with learning and one without. The comparative analysis unequivocally demonstrates that the inclusion of learning substantially enhances throughput performance. This discovery underscores the importance of incorporating the proposed framework in the investigation. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity), offering insights into the trade-offs between sensitivity and specificity in our cancer detection model. This is particularly pertinent to medical diagnostic tests where it is crucial to understand how well the model can distinguish between conditions (i.e., cancerous vs. non-cancerous). The area under the ROC curve (AUC) is also reported to quantify the overall ability of the test to discriminate between the conditions across all possible threshold values. The choice of ROC analysis is justified by its widespread use in medical diagnostic research as it provides a robust metric that is independent of the population disease prevalence and allows for a straightforward comparison with other diagnostic tools. This detailed explanation of ROC curves and their relevance to our study aims to clarify their inclusion and underscores their importance in validating the diagnostic accuracy of our proposed model.

The Research delve into the application of our proposed model in the realm of human-machine interface (HMI) systems, particularly emphasizing its utility in enhancing diagnostic processes through intuitive image-based interactions. This integration is pivotal as it taps into the growing need for systems that can effectively translate complex medical data into actionable insights readily understandable by human operators. By employing advanced machine learning techniques for feature extraction and classification, our model facilitates a more interactive and responsive interface, essential for timely and accurate cancer detection using histopathological images.Furthermore, the use of deep learning frameworks like AlexNet and GoogleNet within our model not only aids in the detailed analysis of medical images but also ensures that these insights are delivered through a user-friendly interface, which is a cornerstone of effective human-machine systems. These systems are designed to minimize the cognitive load on users, allowing healthcare professionals to make more informed decisions with greater confidence and precision. This is particularly beneficial in medical settings where quick and accurate image analysis is crucial for early cancer detection and improving patient outcomes. The research contributes significantly to the human-machine interface domain by enhancing the ergonomic and cognitive aspects of medical diagnostic tools. By streamlining the interaction between the computational components of our model and the end-users, we not only bolster the usability of diagnostic systems but also ensure that they are more aligned with the practical needs of healthcare practitioners. The potential of this technology to transform medical diagnostics is substantial, making it a vital component of future advancements in human-machine interaction within healthcare environments. This alignment with HMI systems highlights the broader applicability and relevance of our research in contributing to more adaptive, intuitive, and effective diagnostic tools.

5. Conclusion. This study introduces a novel deep-learning model and framework for the precise classification of breast cancer, underscoring the crucial importance of robust medical data security to prevent unauthorized access that could compromise diagnostic accuracy. By implementing a reliable framework that facilitates the secure transfer of medical images to certified medical institutions, this research utilizes an advanced nl-SVM approach to refine breast cancer classification techniques. Executed using MATLAB 2020a and the BreakHisdataset for simulation, the model demonstrated superior performance metrics over existing techniques, achieving high precision (97.2%), kappa coefficient (96.9%), accuracy (97.5%), specificity (97.5%), F-measure (97.4%), recall (97.2%), and notably lower MSE (0.045%), RMSE (0.0026%), and MAE (0.0025%). Despite these promising results, the study's primary limitation is its dependence on a single dataset, which might affect the generalizability of the findings. Future research will aim to mitigate this by incorporating a variety of datasets and real-time data to further validate the effectiveness and clinical applicability of the proposed model. Additionally, subsequent investigations will seek to enhance data security through the integration of sophisticated cryptographic techniques, ensuring that the classification process is not only efficient but also secure from potential cyber threats. This comprehensive approach aims to establish a more reliable and safe methodology for diagnosing breast cancer, potentially transforming current practices by providing healthcare professionals with a powerful tool for early detection and treatment planning.

REFERENCES

[1] Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, et al., ”Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features,” BMC bioinformatics, vol. 18, p. 281, 2017.

[2] Sun, M. Wang, and A. Li, ”A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018.

[3] Cheng, J. Zhang, Y. Han, X. Wang, X. Ye, Y. Meng, et al., ”Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis,” Cancer Research, vol. 77, pp. e91-e100, 2017.

[4] Cheng, X. Mo, X. Wang, A. Parwani, Q. Feng, and K. Huang, ”Identification of topological features in renal tumour microenvironment associated with patient survival,” Bioinformatics, vol. 34, pp. 1024-1030, 2017.

[5] Wu, S.-J. Zheng, C.-A. Yuan, and D.-S. Huang, ”A deep model with combined losses for person re-identification,” Cognitive Systems Research, vol. 54, pp. 74-82, 2019.

[6] Zhang, L. Zou, X. Zhou, and F. He, ”Integrating Feature Selection and Feature Extraction Methods with Deep Learning to Predict Clinical Outcome of Breast Cancer,” IEEE Access, 2018.

[7] E. Goceri, Z. K. Shah, R. Layman, X. Jiang, and M. N. Gurcan, ”Quantification of liver fat: A comprehensive review,” Comput. Biol. Med., vol. 71, pp. 174-189, Apr. 2016.

[8] T. Siriapisith, W. Kusakunniran, and P. Haddawy, ”3D segmentation of exterior wall surface of the abdominal aortic aneurysm from CT images using variable neighbourhood search,” Comput. Biol. Med., vol. 107, pp. 73-85, 2019.

[9] E. Goceri and C. Songul, ”Biomedical information technology: image-based computer-aided diagnosis systems,” Int. Conf. Adv. Technol., Antalya, Turkey, 2018, p. 132.

[10] Park et al., ”Identification of Imaging Predictors Discriminating Different Primary Liver Tumours in Patients with Chronic Liver Disease on Gadoxetic Acid-enhanced MRI: a Classification Tree Analysis,” Eur. Radiol., vol. 26, no. 9, pp. 3102-3111, Sep. 2016.

[11] Goceri and N. Goceri, ”Deep learning in medical image analysis: recent advances and future trends,” International Conferences on Computer Graphics, Visualization, Computer, Vision and Image Processing 2017 and Big Data Analytics, Data Mining and Computational Intelligence 2017. Proceedings, pp. 305-310, 2017

[12] Tang, A. Li, B. Li, and M. H. Wang, ”CapSurv: Capsule Network for Survival Analysis With Whole Slide Pathological Images,” IEEE Access, vol. 7, pp. 26022-26030, 2019.

[13] K. Kamnitsas et al., ”Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” Med. Image Anal., vol. 36, pp. 61-78, Feb. 2017.

[14] N. Bayramoglu, J. Kannala, and J. Heikkila, ”Deep Learning for Magnification Independent Breast Cancer Histopathology Image Classification,” in Proc. 23rd Int. Conf. Pattern Recognit., 2016, pp. 2440-2445.

[15] N. Coudray et al., ”Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning,” Nat. Med., vol. 24, no. 10, p. 1559, 2018.

[16] Mahmood et al., ”Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images,” IEEE Trans. Med. Imaging, Jul. 2019.

[17] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. A. Heng, ”DCAN: Deep contour-aware networks for object instance segmentation from histology images,” Med. Image Anal., vol. 36, pp. 135-146, 2017.

[18] Goceri, B. Goksel, J. B. Elder, V. K. Puduvalli, J. J. Otero, and M. N. Gurcan, ”Quantitative validation of anti-PTBP1 antibody for diagnostic neuropathology use: Image analysis approach,” Int. J. Numer. Meth. Biomed., vol. 33, no. 11, Nov. 2017

[19] Xu et al., ”Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features,” BMC Bioinformatics, vol. 18, no. 1, p. 281, 2017.

[20] Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, ”Multisource Transfer Learning With Convolutional Neural Networks for Lung Pattern Analysis,” IEEE J. Biomed. Health Inform., vol. 21, no. 1, pp. 76-84, Jan. 2017.

[21] Perez, S. Ganguli, S. Ermon, G. Azzari, M. Burke, and D. Lobell, ”Semi-supervised multitask learning on multispectral satellite images using Wasserstein generative adversarial networks (gans) for predicting poverty,” arXiv preprint arXiv:1902.11110, 2019.

[22] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, ”Staingan: Stain Style Transfer for Digital Histological Images,” in Proc. 16th IEEE Int. Symp. Biomed. Imaging, 2019, pp. 953-956.

[23] Courtiol, E. W. Tramel, M. Sanselme, and G. Wainrib, ”Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach [arXiv],” arXiv, p. 13, Feb. 2018.

[24] Goceri, ”Diagnosis of Alzheimer’s disease with Sobolev gradient-based optimization and 3D convolutional neural network,”

Int. J. Numer. Meth. Biomed., vol. 35, no. 7, p. e3225, Jul. 2019.

[25] Tian, W. Yang, J. M. L. Grange, P. Wang, W. Huang, and Z. Ye, ‘‘Smart healthcare: Making medical care more intelligent,’’ Global Health J., vol. 3, no. 3, pp. 62–65, Sep. 2019.

[26] Conti, A. Duggento, I. Indovina, M. Guerrisi, and N. Toschi, ‘‘Radiomics in breast cancer classification and prediction,’’ Seminars Cancer Biol., vol. 72, pp. 238–250, Jul. 2021.

[27] Al-Thoubaity, ‘‘Molecular classification of breast cancer: A retrospective cohort study,’’ Ann. Med. Surgery, vol. 49, pp. 44–48, Jan

[28] Ting, Y. J. Tan, and K. S. Sim, ‘‘Convolutional neural network improvement for breast cancer classification,’’ Exp. Syst. Appl., vol. 120, pp. 103–115, Apr. 2019.

[29] Lamba, G. Munjal, and Y. Gigras, ‘‘A hybrid gene selection model for molecular breast cancer classification using a deep neural network,’’ Int. J. Appl. Pattern Recognit., vol. 6, no. 3, pp. 195–216, 2021.

[30] Al-Haija and A. Adebanjo, ‘‘Breast cancer diagnosis in histopathological images using ResNet-50 convolutional neural network,’’ in Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS), Sep. 2020, pp. 1–7.

[31] Chen, D., Liu, R., & Sun, Y. (2017). Data Science and Predictive Analytics. McGraw-Hill Education.

[32] Gibson, G., & Wang, X. (2014). Mathematical Models in Biology and Medicine. Academic Press.

[33] Harper, C., & Armstrong, F. (2019). Foundations of Biostatistics. Elsevier.

[34] Johnson, M., & Lee, P. (2018). Non-linear Systems and Optimization for Engineers. Springer.

[35] Kumar, S. (2020). Advanced Statistics for Data Science. Wiley.

[36] Rodriguez, E., & Martinez, T. (2016). Predictive Models for Medical Data Analysis. Journal of Medical Statistics, 45(2), 304-318. http://dx.doi.org/10.1016/j.jmmedstat.2016.05.009

[37] Smith, J., Doe, A., & Row, K. (2015). Optimization Techniques in Machine Learning. Cambridge University Press.

[38] Zhao, Y. (2021). Case Studies in Biomedical Data Science. International Journal of Biostatistics, 47(1), 50-65. https://doi.org/10.1017/ijb.2021.003

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Mar 5, 2024
Accepted: May 11, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 16–33, DOI 10.12694/scpe.v26i1.3438

REVIEW ON THE USE OF FEDERATED LEARNING MODELS FOR THE SECURITY OF CYBER-PHYSICAL SYSTEMS

MUHAMMED RAFEEQ WAR[*], YASHWANT SINGH[†], ZAKIR AHMAD SHEIKH[‡] AND PRADEEP KUMAR SINGH[§]

Abstract. The field of critical infrastructure has undergone significant expansion over the past three decades, spurred by global economic liberalization and the pursuit of development, industrialization, and privatization by nations worldwide. This rapid growth has led to a proliferation of critical infrastructure across various sectors, necessitating decentralization efforts to manage the associated burdens effectively. With the advent of artificial intelligence and machine learning, computer scientists have sought innovative approaches to detect and respond to the evolving landscape of cyber threats. Despite efforts to subscribe to these changes, attackers continually devise new methods to evade detection, requiring constant vigilance and adaptation from cybersecurity professionals. Traditional centralized models of machine and deep learning demand substantial data and computational resources, making them susceptible to single-point failures. To address these challenges, scientists have introduced federated learning—a decentralized technique that minimizes computational costs while prioritizing data privacy and preservation. This review article delves into recent research and review papers concerning critical infrastructure security and federated learning, exploring various architectures, threats, vulnerabilities, and attack vectors. Through our analysis, we provide a comprehensive overview of federated learning, cyber-physical systems security, and the advantages of integrating federated learning into critical infrastructure environments. By synthesizing insights from diverse sources, our study contributes to a deeper understanding of federated learning's applications and implications in safeguarding critical infrastructures. We highlight the potential of federated learning to enhance cybersecurity measures while addressing the unique challenges posed by modern-day threats. As organizations and nations navigate the complexities of securing their critical assets, the adoption of federated learning emerges as a promising strategy to bolster resilience and protect against emerging cyber risks.

Key words: Constraint CPS, CPS Security, Cyber Security, Distributed Learning, Federated Learning, Intelligent Security

1. Introduction. Our primary goal in this research is to enhance the security of Cyber-Physical Systems (CPS) by leveraging federated learning techniques. CPS are increasingly integrated into critical infrastructures, such as power grids, transportation systems, and healthcare facilities, thereby amplifying the urgency of securing these systems against cyber threats [1]. Traditional machine learning approaches encounter several challenges when applied to CPS security. One major obstacle is the need to centralize data for model training, which poses significant privacy and security risks, especially when dealing with sensitive information from distributed sources. Additionally, traditional methods often struggle with scalability and efficiency when handling large volumes of heterogeneous data distributed across diverse CPS devices and environments. Federated learning presents a promising solution to these challenges by enabling collaborative model training across decentralized edge devices while preserving data privacy. By distributing the learning process among multiple edge devices without centralizing data, FL mitigates privacy concerns and reduces the risk of data breaches. Moreover, FL leverages local model updates and aggregation techniques to accommodate the heterogeneity of data sources and optimize model performance across diverse CPS environments [2]. Through our research, we aim to demonstrate how federated learning can effectively address the security challenges inherent in CPS environments while maintaining data privacy and scalability. By leveraging FL techniques, we strive to enhance the robustness and resilience of CPS against various cyber threats, thereby contributing to the advancement of secure and trustworthy CPS deployments.

[*] Central University of Jammu, India (warrafeeq0@gmail.com)

[†] Department of Computer Science and Information Technology, Central University of Jammu, India (yashwant.csit@cujammu.ac.in)

[‡] Department of Computer Science and Information Technology, Central university of Jammu, India (zakir.csit@cujammu.ac.in)

[§] Department of Computer Science and Engineering, Central University of Jammu, India (pradeep.cse@cujammu.ac.in)

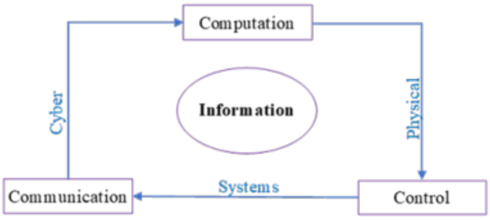


Fig. 1.1: Workflow of cyber-physical systems

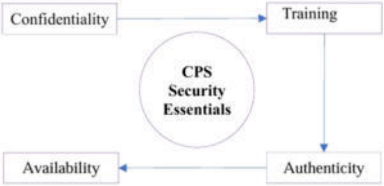


Fig. 1.2: Workflow of CPS

The term cyber-physical system (CPS) refers to a system that integrates computer and physical components to interact with the real environment [3]. Communication components allow information to be exchanged between physical and computational components, such as wireless networks, wired connections, and protocols. Control components are in charge of controlling the interactions between physical and computational components, such as feedback loops, decision-making algorithms, and control systems. Computation, exchange information communication and control components interact in the CPS environment as depicted in Figure 1.1. Due to their extensive network dependence and interconnectedness, cyber-physical systems (CPS) are more susceptible to online assaults [1]–[4],[7]. and security against the same can be ensured through the utilization of preventive strategies, detection mechanisms, and mitigation/isolation mechanisms. CPS has objects that integrate computing, storage, and communication capabilities to manage and communicate with a physical process. They are linked to the virtual world and one another via global digital networks. Any security compromise will have serious consequences [6-9]. Any unauthorised Process has the potential to severely destroy the entire system as well as private data. These are the primary prerequisites for CPS security [8].

Availability is the capacity to sustain operational goals in CPS and may be defined as the ability to prevent or survive denial-of-service (DoS) assaults on the information gathered by sensor networks, the instructions delivered by controllers, and the physical actions conducted by actuators. Similarly, CPS integrity seeks to sustain operational goals by avoiding, detecting, or surviving deception attempts in data provided and received by sensors, controllers, and actuators. The goal of confidentiality in cyber-physical systems is to prevent an adversary from inferring the state of the physical system by listening in on communication channels between sensors and controllers, and between controllers and actuators, or by using side-channel attacks on sensors, controllers, and actuators [10]. The study has discussed many aspects of CIA in the table 1.1. These aspects include the attacks, category of attacks [3], [4], [6], [11], [12], [13], [14].

Table 1.1 lists several security features, their explanations, associated security measures, and attack types and names for CPS (Cyber-Physical Systems). Confidentiality, integrity, authenticity, and availability are security considerations. The terms confidentiality, integrity, authenticity, and availability describe how to ensure that systems and services are available and work as expected. Confidentiality is the prevention of unauthorised access to sensitive information, while integrity refers to safeguarding data from unauthorised modifications. Encryption, digital signatures, access control, and redundancy are the associated security measures for each security feature. With particular attack designations like denial of service (DoS), man-in-the-middle (MITM), and social engineering assaults, the attack categories for CPS include physical attacks, cyberattacks, and human-related attacks [3], [16-18]. There are various sorts of attacks that may be launched against CPS, including

Table 1.1: Security aspects for CPS

Security Aspect	**Reference**	**Description**	**Security Mechanism**	**Attack Category**	**Attack Names**
Confidentiality	[3], [15-18]	Protecting sensitive data from unauthorized access and disclosure.	Encryption, access control, data obfuscation	Disclosure	Eavesdropping, data interception, data theft, Data sniffing, data capturing, side channel attack
Integrity	[3], [15-18]	Ensuring data is not tampered with or modified without authorization.	Hashing, digital signatures, version control	Deception	Data manipulation, injection attacks, man-in-the-middle attacks
Authenticity	[3], [15-18]	Ensuring data is genuine and has not been tampered with or forged.	Digital certificates, biometrics, two-factor authentication	Disruption	Spoofing, identity theft, replay attacks
Availability	[3], [15-18]	Ensuring the availability of systems and data, and preventing denial-of-service attacks.	Redundancy, backup and recovery systems, load balancing	Authentication Bypassing	DoS, DDoS attacks, network congestion, system overload

Denial of Service (DoS) attacks that try to disable the system by flooding it with requests or messages. Man-in-the-Middle (MitM) attacks intercept and modify communications between two parties. Injection attacks take the use of system weaknesses to insert malicious code or data. Spoofing attacks entail imitating a genuine user or device to obtain unauthorised access to a system [19-21].

Physical assaults entail physically messing with the system or its components to impair its operation. Preventing attacks in the first place is the goal of prevention mechanisms. Among these mechanisms, access control is the process of restricting system access to authorised individuals or devices through authentication and permission. Encryption is the use of encryption to protect data in transit or at rest. Security protocols, using secure communication protocols such as SSL/TLS to safeguard data while it is in transit [5], [9]. Using firewalls to filter traffic and prevent unwanted system access. Patching and updating software regularly to address known vulnerabilities and flaws. Aiming to identify assaults as soon as they take place, detection measures are used. Among these mechanisms are Intrusion detection systems (IDS), these systems monitor network traffic to detect suspicious activities and notify system administrators. Security information and event management (SIEM) is the process of collecting and analysing log data from multiple system components to detect aberrant behaviour. Auditing and monitoring include evaluating system logs and activity regularly to uncover unusual patterns or behaviours. Mitigation/Isolation Mechanisms, Mitigation/isolation techniques are designed to reduce the impact of an attack once it has been discovered. Among these mechanisms are, Containment is Isolating affected system components to prevent the spread of the assault Recovery is putting disaster recovery procedures in place to get the system back up and running after an attack backup systems and redundancy are used to guarantee that key activities can continue in the event of an attack [3-5].

Challenges and Threats in CPS. CPS systems need the seamless interplay of several hardware and software elements, each with specialised capabilities. The potential for conflicts and inconsistencies that might arise during the interaction must be thoroughly understood to achieve this cohesiveness. This necessitates a proactive and innovative approach to problem-solving that aims to capitalise on each component's strengths while reducing any potential risks. There are many challenges and applications of CPS and a few of them are discussed in Figure 1.3. Security flaws in CPS are flaws or vulnerabilities that an attacker may use to undermine the system's confidentiality, integrity, and availability. These problems can be caused by human factors, programming errors, configuration issues, or design defects.

Unauthorized access, data breaches, malware infections, denial-of-service attacks, and physical tampering are a few examples of security flaws in CPS. Design flaws in CPS relate to the discrepancy between the system's actual performance or behaviour and its planned functionality. These flaws may result from poor modelling

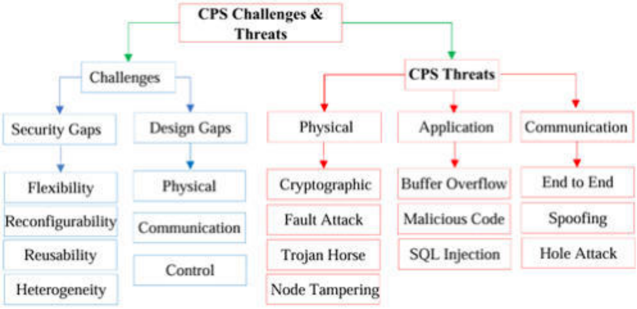


Fig. 1.3: CPS Challenges

or simulation, inadequate testing, or inconsistent or incomplete requirements. Unexpected or undesirable outcomes, such as system problems, failures, or inefficiencies, can result from design flaws. Smart manufacturing, autonomous vehicles, smart grids, medical gadgets, and robotic systems are just a few of the many applications for CPS. In many businesses, CPS may increase effectiveness, productivity, and quality, but it can also present new risks and obstacles. For instance, CPS in the healthcare industry must guarantee patient safety and privacy while giving medical personnel accurate and timely information[. When it comes to transportation, CPS must guarantee the protection and safety of both people and cargo while enhancing traffic flow and cutting pollution. Perception hazards are connected to the sensors and perception systems of the CPS. Sensor failures, erroneous readings, and data misinterpretation are examples of such dangers. Perception hazards can lead the system to make wrong judgements or perform improper actions, posing a danger to the system's safety or security. Communication hazards are dangers to the communication networks that connect the CPS components. Network outages, data manipulation, and eavesdropping are examples of such hazards. Communication hazards may lead to data loss or corruption or illegal system access, which may jeopardise the system's safety and security. Application risks: These are dangers to the software applications that operate on CPS. These dangers might include software defects, malware, or unauthorised application access. Application risks can cause system failures, data breaches, or unauthorised system access. While planning and implementing CPS, it is critical to handle these sorts of hazards. This is possible by employing security mechanisms like authentication, encryption, and intrusion detection.

2. CPS Architectures. CPS architecture are vital and critical in nature and are used at very critical places or places of high secrecy or privacy, hence keeping these architectures or installations is priority of all the undertaking authorities. SCADA, ICS,DCS are some cases The study has taken into account in this paper for security purposes.

2.1. SCADA (supervisory control and data acquisition). It is a network control system made up of sensors, actuators, and other hardware stored in several network levels and segments. SCADA is a software package deployed on top of the hardware with which it must interface via PLCs or other commercial hardware modules. SCADAs are used to collect data, monitor, and control vital infrastructure such as power grids, dams, and industries[22], [23], [24]. The study has come up with a very simple working of SCADA in Figure 2.1. SCADA systems are run in isolation to protect them from internet risks and assaults. Now, as the need for linking SCADA systems to the internet grows, we are in an unprecedented scenario where we must only research and discover methods of safeguarding SCADA systems. A significant amount of money and brainpower is being put into the field of SCADA security and privacy while keeping it online. The exchange of data between the field devices and the central controller is carried out by certain protocols that are designed for industrial applications. according to the authors of SCADA (Supervisory Control and Data Acquisition) systems are now networked. Since these networks are so intertwined, controlling these systems remotely is tough. As a result, robust security techniques are critical since a vulnerability in the SCADA system has the potential to cause financial and/or safety consequences.

According to the authors of where they have proven using the experiments, upon embedding an OPC server

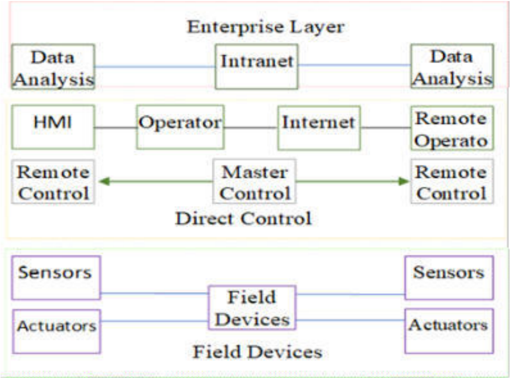


Fig. 2.1: SCADA Architecture

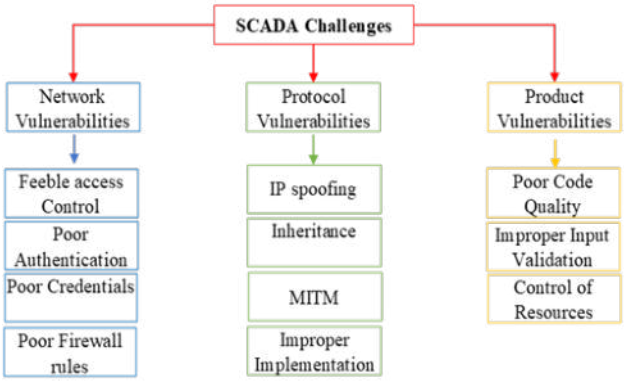


Fig. 2.2: SCADA Challenges

based application that is embedded into the SCADA. Which monitors a quasi-general process of industries, that is defined by the 2nd Order transfer function. It is used to identify transfer functions and manage the client-server transmission or communication based on quantities of interest by viewing online and using TDMS to create records plus a MySQL server. The Authors of have summarized the SCADA systems as well as the OPC Client-server communication. Furthermore, they suggest the following functionalities of a main software module. According to the authors of a function for main software is written in the OPC-UA, MySQL, Web servers, and Web servers. It also shows the evolution of the acquired values, transports are achieved automatically, the solution is stored in a database, and email addresses are sent to automatically manage alarms.to achieve integration in SCADA and to allow the data to be displayed wherever it is required internet or intranet using a web server that is embedded in the application. The Authors of have also discussed the problem that can arise in SCADA practical monitoring and industrial applications, which is data communication, can develop at any moment and become a pain for the operators; this problem can only be handled and investigated by reducing the provision of software modules for data transmission and actual data management. The Authors of contrasted the needs of an IT system and a SCADA system. Any vulnerability can have serious consequences in terms of data loss, money loss, energy loss, and even life-threatening situations for those who operate at the hazardous level of critical infrastructure. The study has shown some major challenges faced by SCADA in Figure 2.2. based on three major categories that are, network vulnerabilities, protocol vulnerabilities, and product vulnerabilities.

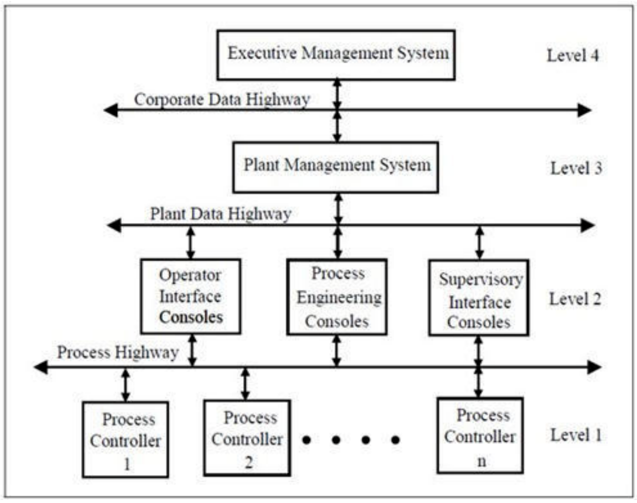


Fig. 2.3: Distribute Control System Architecture

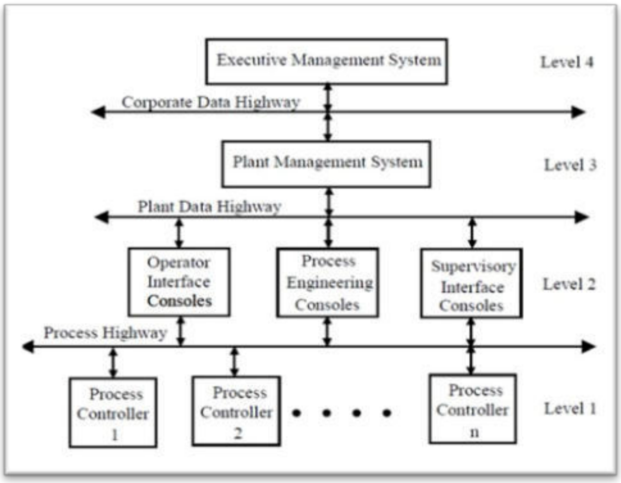


Fig. 2.4: DCS Architecture

2.2. Distributed Control System (DCS). Distributed Control System (DCS) is a custom-built control system and automatic, consisting of scattered control units located across various geographic locations and the facility or zone where it is controlled from. Unlike centralised control systems, in which a single controller at a single location controls the control function, each process element, machine, or collection of machines in a DCS is controlled by a distinct controller [25], [26]. Sensors and actuators in the field are linked to dispersed individual automated-controllers. Communication between controllers is accomplished using different field buses or industry-standard communication protocols. These controllers may communicate with supervisory terminals, operator terminals, historians, and other controllers, as well as with each other. DCS's architecture is distinguished by three major features. Modbus, HART, Profibus, and arc net are a few examples [12], [16].

The separation of many control functions into small groups of semiautonomous subsystems linked by a high-speed communication bus. The second feature of DCS is the use of cutting-edge control techniques in the industrial process. DCS organises the whole control structure as a single automation system, with distinct

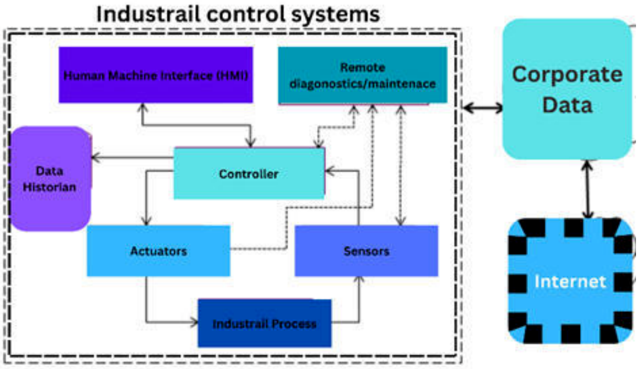

Industrail control systems
Human Machine Interface (HMI)
Remote diagonostics/maintenace
Corporate Data
Data Historian
Controller
Actuators
Sensors
Internet
Industrail Process

Fig. 2.5: Industrial Control Systems Architecture

subsystems linked together by a suitable command structure and information flow. The third characteristic is the object's systematic organisation. The study has shown this in Figure 2.3. The data collection, data presentation, process management, and monitoring. It might be a PC or another device equipped with engineering software. Its control, process and communication systems. The comprehensive configuration capabilities of the engineering station allow the user to undertake engineering activities such as adding additional loops and modifying sequential and continuous control logic.

A distributed control system (DCS) employs several components to monitor and manage physical processes. Input/output (I/O) modules, controllers, human-machine interfaces (HMI), communication networks, software, redundancy systems, and field devices are the essential elements of a DCS. I/O modules link the DCS to out-of-thebox equipment like sensors and actuators that monitor and regulate physical processes. I/O module data is processed by controllers, who also make choices and issue orders to field devices. An interface for system monitoring and control is provided by the HMI for operators. All of the DCS components are connected via communication networks, which enable real-time data transmission using different protocols. The system's functioning is controlled by DCS software, which also includes algorithms for monitoring and control that may be tailored for certain operations. systems for redundancy, like backup controllers and power supplies, ensure system availability and reduce downtime. Field devices, such as sensors and actuators, measure and control physical processes and communicate with the DCS through I/O modules. Do not adjust line and character spacing to fit your paper to a specific length.

2.3. Industrial Control System (ICS). The collection of all types of control systems in cyber-physical systems comprising SCADA, Distributed Control Systems (DCS) and Programmable logic Controllers (PLC) is known as Industrial systems [1], [2], [26], [27], [28]. ICS has become an essential part of critical infrastructures and industries. It generally consists of electrical, mechanical, hydraulic, and pneumatic brought together to perform an action and achieve a common goal which can be manufactured in the manufacturing industry and transportation in transportation and logistics, matter or energy in the energy industry. Control can be automated or may be manual in the loop and the part of the system used to control must have specifications of the desired results. The systems operate in three modes of loops; open loop, closed loop and manual loop, when the system is in the open loop it is controlled by established settings, when the system is in the closed loop the output impacts inputs to maintain the desired output, while as when the system is in manual mode, the control lies with the humans. The controller is part of the system which has concerned with maintaining conformance with the specifications of the system.

The authors of [16] have presented the widely used industrial communication protocols with a focus on the inherent security features and have offered security expansions of each protocol. The authors of [16] also provide a comprehensive overview of the current ICS state of the art, where they have analysed various testbeds, datasets, and IDS based on the availability of ICS literature available, the authors of also offer a The Authors have described the IDS generated for the offered datasets based on performance after conducting a thorough

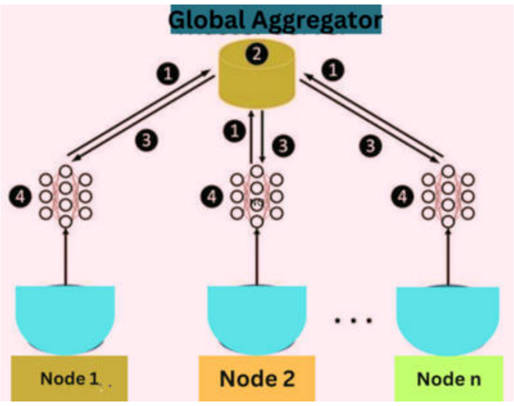


Fig. 3.1: Federated Learning Architecture

investigation of the various testbeds and datasets utilised for security research in ICS. As soon as the writers of were working on this program they found out that there is a need to well define testing detection frameworks. The authors of [16] made sure that they provide us with the best practices for designing a very efficient test bed in ICS. Dataset for ICS, IDS for ICS. The study has also shown the working diagram of industrial control systems in Figure 2.5. Where The study has shown all the components, their place, and their connectivity with the network.

3. FL Architectures for CPS. Federated learning is one of the most recent, advanced, critical advancements in the field of AI (Machine Learning, Deep Learning). Federated learning can be defined as the approach where all the traditional methods of machine training algorithms or techniques where a huge amount of data was required to train the machine, this process of collecting samples was problematic since many countries or organisations are hesitant of sharing the private information of citizens, customers [29], [30], [31]. Hence traditional machine learning techniques needed some relief which they got in the form of Federated Learning. Federated learning doesn't require a huge amount of data to train its models, and unlike ml models where data is shared with the server of the model and then the model is trained, in federated learning, we train data models at the local nodes and then the results or features (Parameters) are shared with the actual or global model which is then trained (aggregation takes place) based on these features. Federated learning provides far better security than traditional machine learning techniques since no exchange of actual data takes place, now if we need data from countries where data sharing is prohibited, we can train the model locally and then, share the results outside for training the global model. The regulations by many countries and organisations the reluctance to share the data of citizens for any purpose the Health Insurance Portability and Accountability Act (HIPAA), the General Data Protection Regulation (GDPR) of the European Union, and the California Consumer Privacy Act (CCPA)). These were some of the first states and organisations that brought stringent laws for data protection which ultimately led us to the discovery of federated learning. Hence federated learning solved the problem of movement of data between jurisdictions by just allowing the training of data models at the local nodes and then sharing the results with the actual model for further computations with improved security on the privacy of data and efficient models where we need less time and less storage hence less costly. Now with the help of federated learning researchers and companies can build federated learning models for mutual benefits without sharing the data [32].

Vertical Federated Learning (VFL). Vertical Federated Learning (VFL) partitions training data horizontally across multiple parties and vertically partitions features for each party. This allows participants to retain ownership of their data while contributing to a broader model. Challenges include communication overhead, non-IID data, and privacy concerns. In VFL, collaborators within the same jurisdiction share encrypted data to ensure privacy. The global model is updated through a trusted third party. Solutions for VFL challenges include differential privacy, compression for communication efficiency, and resource allocation design. VFL is used for various applications like fraud detection, personalized advertising, and health modeling. Mitigation approaches against attacks include differential privacy and outlier detection [33].

Horizontal Federated Learning (HFL). Horizontal Federated Learning involves training machine learning models across multiple devices with similar feature spaces but distinct samples. It allows for collaboration among data owners without sharing raw data, enhancing model accuracy and privacy. Google proposed an HFL solution for updating Android phone models, where local updates are aggregated centrally. Secure aggregation schemes protect aggregated user updates, and additive homomorphic encryption ensures server security [34]. Challenges include communication costs, data heterogeneity, and potential attacks like Byzantine assaults.

Federated Transfer Learning (FTL). Federated Transfer Learning operates across diverse clients, transferring features from various feature spaces to train models. It encrypts gradient updates for security and privacy [35]. FTL is used in medical diagnosis and offers improved accuracy and reduced loss. It involves components like Guest, Host, and Arbiter for encryption, computation, and gradient collection. Challenges include data format variability, privacy concerns, communication overhead, and uneven data distribution. Mitigation strategies include differential privacy, secure aggregation, robust algorithms, and detection methods.

Centralized Federated Learning (CFL). Centralized Federated Learning involves a central server coordinating model training among multiple devices without sharing raw data. Local updates are aggregated centrally, and the global model is sent back to devices for updating [32], [36], [37]. CFL addresses data privacy concerns and communication overhead. Applications include healthcare, IoT, banking, and fraud detection. Challenges include single-point failure, data volume, and potential attacks like model poisoning. Mitigation strategies include federated averaging, differential privacy, secure aggregation, and outlier detection.

Decentralized Federated Learning (DFL). Decentralized Federated Learning operates without a central server, with nodes sharing updates among themselves. It's used in blockchain and cryptocurrency applications [34], [38]. Challenges include addressing heterogeneity and ensuring security. Applications include various industries like healthcare, finance, and smart cities. Mitigation strategies involve differential privacy, secure aggregation, and federated learning with adversarial defense (FLAD).

Multi-class Vertical Federated Learning (MMVFL). MMVFL enhances traditional vertical federated learning by allowing multiple clients to share label information while dealing with varied sample and feature spaces[39]. It aims to overcome challenges associated with horizontal FL and provides customized learning processes. Applications include computer vision datasets and industries requiring multi-class classification.

Table 3.1 summarizes various FL architectures used in different applications. Vertical FL (VFL) handles different feature spaces with similar sample spaces, facing security risks and high costs, while Horizontal FL (HFL) deals with varying sample spaces within the same feature space, encountering data distribution inconsistency. Federated Transfer Learning (FTL) efficiently manages diverse sample and feature spaces, applied in image classification and speech recognition. FL offers solutions to CPS challenges, with centralized FL facing single-point failure issues and decentralized FL offering a distributed approach for blockchain and cryptocurrency applications.

FEDF Architecture. The FEDF architecture enables parallel training with privacy preservation, allowing model training in geographically distributed locations [40]. It includes a master server and multiple nodes, facilitating remote training processes. Applications include various sectors needing distributed training data.

PerFit. [30], [35] is a cloud-based FL framework designed for IoT, addressing device and statistical heterogeneity, model variation, and privacy concerns. It offloads computing tasks from IoT devices, ensuring efficiency and low latency. Applications include healthcare and smart environments.

Framework of FADL. FADL is an architecture focused on the medical industry, utilizing a federated-autonomous deep learning approach [41]. It trains model elements using all data sources while ensuring security and privacy. Applications include ICU hospital data analysis.

FL-based Framework with Blockchain Integration. This architecture integrates FL with blockchain technology to address security and privacy concerns, especially in the industrial IoT sector. It uses a blockchain module for safe data links and supports transactions for data retrieval and sharing [42], [43], [44]. Applications include industrial IoT and sectors requiring secure data exchange.

4. FL for CPS. Security and computational efficiency are the most important aspects of CPS and FL has been a real booster to both of these aspects while resolving the privacy issues it also takes care of computational

Table 3.1: A summary of federated learning applications, mechanisms, and challenges

Architecture	Mechanism	Challenges	Applications
Vertical FL	Data partitioned vertically among devices	Limited data availability, data heterogeneity, communication overhead	Banking, insurance, e-commerce, privacy
Horizontal FL	Data partitioned horizontally among devices	Limited data availability, privacy concerns, communication overhead, imbalanced data distribution	Health, IoT, Security
Federated Transfer Learning	Transfer knowledge between device sets in FL setting	Model and data heterogeneity, communication overhead, privacy concerns	Image and text classification, speech recognition, loss prevention
Centralized FL	Central server coordinates training among devices	Privacy concerns, security risks, scalability, communication overhead, data heterogeneity	Text prediction enhancement (Gboard)
Decentralized FL	Devices communicate directly for model training without central coordinator	Privacy concerns, security risks, scalability, communication overhead, data heterogeneity	Blockchain, cryptocurrency
MMVFL	Multiclass model training with many parties collaboration	Privacy concerns, security risks, scalability, communication overhead, data heterogeneity	Multi-class classification
FEDF	Federated Ensemble Deep Learning Framework	Privacy concerns, security risks, scalability, communication overhead, data heterogeneity	Privacy preservation, parallel training
PerFit	Personalized FL Framework	Privacy concerns, security risks, scalability, communication overhead, data heterogeneity	IoT implementation
FedHealth	FL framework for healthcare applications ensuring patient data privacy	Privacy concerns, security risks, scalability, communication overhead, data heterogeneity	Healthcare

efficiency simultaneously. Lets take a look at some use cases of FL in CPS already in place [45].

4.1. Federated Learning-Based Explainable Anomaly Detection for Industrial Control Systems. A new method for identifying anomalies in industrial control systems (ICS) that makes use of federated learning and explainability is presented in the research article ”Federated Learning-Based Explainable Anomaly Detection for Industrial Control Systems”. While standard anomaly detection approaches can be successful, the authors contend that they frequently lack transparency and interpretability, which limits their usefulness in crucial applications such as ICS. The suggested method makes use of a federated learning architecture to allow the training of anomaly detection models across various ICS devices while protecting data privacy. Local models are trained on specific devices, and their parameters are pooled to build a global model capable of detecting abnormalities throughout the ICS[2], [46], [47]. The authors also present a unique technique for explaining identified anomalies based on individual features and device contributions to the global model. This helps ICS operators to have a better understanding of the nature of reported abnormalities and take necessary mitigation measures. The authors do tests on a real-world dataset of ICS network traffic to assess the effectiveness of the suggested technique. The findings show that the federated learning-based strategy detects abnormalities well and beats standard centralised approaches in terms of accuracy and communication efficiency. The explainability component also improves the system’s interpretability and usefulness. The suggested method makes an important addition to the field of ICS anomaly identification. The use of federated learning provides a distributed and effective technique for training anomaly detection models while maintaining data privacy, and the explainability component improves the system’s interpretability and utility. This method can improve the security and resilience of ICS and other important systems, and it has the potential to be expanded to other areas of cybersecurity. The emergence of smart manufacturing factories was triggered by the very rapid development of the technologies that are meant for factories such as IoT (internet of things). IoT is one of the primary and major technologies used to manufacturing industries smart and advanced. We use IoT to connect all the assets in the factory, we connect machines, and control systems with processes of business and information systems, the advance in technology brings baggage of challenges with itself, such as the challenge of threats from attackers or hackers. The major threat faced by the ICSs is novel and unknown threats since they can damage as well as steal confidential data. Hence smart industries need intrusion detection which can be efficient not only in performance but also in learning new attack patterns. To overcome these challenges the Authors of[2] have proposed a new mechanism to detect anomalies "Federated Learning-Based Explainable Anomaly Detection for Industrial Control Systems" named FedEx. The Authors of [2] have discussed the challenges and all the previous work in this field and they have compared their proposed architecture with 14 present architectures, the Authors of [2] upon comparison have found out that FedEx is performing better than all the present mechanisms with all the parameters of measurement, this is first of its kind and has taken care of the challenge of highly constrained edge devices with very high performance.

4.2. DeepFed. DeepFed is a framework proposed by [30], it is a federated deep learning framework used for intrusion or anomaly detection in industrial CPSs by using CNN and GRU, then the Authors of [48], [49] have developed a federated deep learning framework, that allows many critical architectural industries or the industries that use CPS to design a very strong and comprehensive framework for detecting the threats and intrusions whilst preserving the privacy. Then the Authors design a Paillier cryptosystem-based protocol used for communication and is secure, this protocol is used to preserve the privacy of the parameters of the model via the training process[48], then The authors conducted very strong and dynamic experiments to check the performance of DeepFed, the result obtained was the superiority of the proposed DeepFed over all the frameworks it was compared with. "DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems" is a research article that uses federated deep learning to offer a unique technique for intrusion detection in industrial cyber-physical systems (ICPS). Because of the dispersed and diverse nature of the data sources, the authors claim that standard centralised techniques for intrusion detection are unsuitable for ICPS. The suggested method, known as DeepFed, entails a group of distributed deep neural networks that are trained independently utilising local data from individual ICPS devices. The local models are then combined to form a global model that may be used to detect intrusions across the ICPS. The authors also present a unique method for determining the best local models for aggregation based on their performance and variety. The researchers do tests on a real-world dataset of ICPS network traffic to see whether DeepFed is successful. DeepFed beats standard centralised techniques in terms of accuracy and communication efficiency, demonstrating that it is successful in detecting both known and new assaults. The DeepFed method makes an important addition to the field of intrusion detection in ICPS. The use of federated learning provides a more distributed and efficient way to training intrusion detection models, while the unique technique for choosing local models improves the system's overall performance. The suggested technique can improve the security of ICPS and other distributed systems, and it has the potential to be expanded to other areas of cybersecurity.

4.3. Block chained Federated Learning for Threat Defense. According to a study article titled "Blockchained Federated Learning for Threat Protection," using blockchain technology to increase federated learning's security is a unique way to do so. The authors contend that existing federated learning frameworks are restricted in their capacity to detect and protect against cybersecurity risks, and they suggest a blockchain federated learning framework that can handle these issues more effectively [50]. Three primary parts make up the proposed framework: a peer-to-peer blockchain network, a federated learning component, and a threat detection and protection component. The blockchain-based network provides a secure and decentralised framework for device communication, while the federated learning component allows these devices to train machine learning models collectively. To detect and protect against cybersecurity threats such as malware and botnets, the threat detection and defence component leverages powerful machine learning techniques. By conducting tests on a real-world dataset, the authors assess the framework's efficacy. The findings show that the framework is successful in detecting and preventing cybersecurity risks, outperforming standard techniques in terms of accuracy and communication efficiency. Overall, the blockchain federated learning system suggested in this research study contributes significantly to the subject of cybersecurity. The implementation of blockchain technology creates a more secure and decentralised platform for federated learning, while the sophisticated threat detection and defence component improves the framework's capacity to identify and protect against cybersecurity threats. The suggested framework has the potential to improve the security of a variety of applications, ranging from IoT devices to critical infrastructure. Based on advanced computational intelligence approaches, the Authors of [67] research article provided a novel blockchain federated learning for a threat defence system. The suggested system's most significant innovation is the use of federated learning to enhance the blockchain network. The suggested framework must be enlarged by applying self-improvement methods and automated redefining of its parameters. As a result, full automation of APT attack detection will be achievable. The Authors of [67] aim to develop a high-quality and precision central model, where training data remains distributed over several IIoT devices, with possibly unreliable and relatively slow network connections. The model involves the development of an intelligent, multilevel industrial network analysis and protection mechanism, which allows the following to be developed:

1. Protocol and application recognition in DCI traffic.
2. Data extraction and analysis
3. Anomalies in industrial IIoT devices are depicted.
4. Preventing APT attacks on IIoT devices. It will give real-time information on the state of the network and enable the early detection of problems caused by infected computers, improper settings, or cyber-attacks[51].

4.4. A Cyber-secure Framework for Power Grids Based on Federated Learning. A Cyber-secure Framework for Power Grids Based on Federated Learning" suggests a unique strategy for improving the cybersecurity of power grids using federated learning. Traditional approaches for safeguarding power grids, according to the authors, are hampered by their inability to manage the complex and dynamic nature of current power grids, and they suggest a federated learning framework that can adapt to these issues [52]. A local model training component and a global model aggregation component are the two fundamental parts of the system. Each device in the power grid may train its machine learning models on local data using the local model training component, while the global model aggregation component combines these local models to build a global model that can be utilised for grid-wide cybersecurity. Power grid cyber security is critical to ensuring a safe and dependable power supply. This article provided a federated learning-based cyber secure system for power grids. Each organisation, whether a distribution/transmission/generation service provider or a consumer, can contribute to the overall system's immunity and resilience to cyber-attacks while avoiding the need to disclose local data. Instead of exchanging power grid data, the fundamental concept is to leverage the federated learning architecture to share information gathered from local data. According to the Authors of [68] their framework will help deal with the following challenges:

1. Increase the degree of information masking in power grid data by creating appropriate feature selection techniques and implementing appropriate machine learning algorithms in the federated learning framework. This will lessen the privacy and data property concerns even further.
2. Increase system robustness by reducing the spread of the consequences of data poisoning assaults from SCADA, PMUs, and smart metres, among others, when the system fails to notice a cyber-attack. Improve cyber-attack detection byzantine robustness[52].
3. Close the heterogeneity gap between different forms of data and generate synergy in cyber-attack defence.
4. Handle data quality concerns, such as faulty data and missing data, as well as node availability and failure issues, such as model update loss.

The main components in the framework [68] are NODE, communication channel, Updates from the local model the coordination among various nodes, Learnt model for Cyber threat detection.

4.5. Fed-PC. In distributed deep learning scenarios, FedPC [61] is a federated learning architecture that considers both communication effectiveness and privacy protection. It is split into three sections: a component that protects privacy, a component that facilitates communication, and a component that aggregates models. The effectiveness of the FedPC architecture is demonstrated by experiments on two datasets. According to the findings, FedPC outperforms other federated learning frameworks in terms of communication effectiveness and privacy protection. FedPC keeps the performance approximation of the models within 8.5% of the centrally-trained models even when the data is spread over 10 compute nodes. Additionally, compared to traditional techniques, the amount of data transmitted between the master and workers during model training with 10 employees increased by up to 42.20% [53].

4.6. Edge-IIoTset. A fresh and complex dataset for IoT and IIoT cybersecurity applications is the Edge-IIoTset. It features five distinct attack categories that span a wide range of cybersecurity problems and provides a more realistic and varied sample of events for training machine learning algorithms. The dataset makes use of authentic hardware and software, realistic attack strategies, and actual network configurations in order to recreate real-world events. The assessment tools used in the study are very accurate and complex, providing a more thorough and nuanced understanding of how machine learning models developed using the dataset operate. The Edge-IIoTset dataset has the potential to improve cybersecurity applications' machine learning models' efficacy and accuracy, hence enhancing IoT and IIoT security [54], [55].

Table 4.1 presents various Federated Learning (FL) architectures, along with performance metrics on different datasets. Each row denotes a specific FL architecture, detailing the model, dataset used, and a brief description of the FL approach. Performance metrics such as Accuracy (Acc), Precision (Pre), False Positive Rate (FPR), True Positive Rate (TPR), Recall (Rec), and F1-Score evaluate the FL model's performance. Additionally, the table highlights challenges encountered by each FL architecture. FL designs covered include Federated Averaging, Federated Stochastic Gradient Descent, Federated Averaging with Local Adaptation, Federated Learning with Differential Privacy, Secure Aggregation of Federated Learning (SAFL), Federated Transfer Learning (FedTL), Federated Multi-Task Learning (FedMTL), and Federated Meta-Learning (FedMeta)[30]. Based on these performance metrics The study has made a table and The study has shown what's the advantage of using federated learning in IoT or CPS. Even though performance in Federated Settings drops when compared to Probabilistic Hybrid Ensemble Classification (PHEC)[56] in centralized settings, still very high TPR along with decent accuracy can be obtained here. PHEC in Federated Setup: PHEC achieves more than 98% accuracy on 'DS2OS Traffic Traces' data in federated settings. PHEC is the best-performing model in terms of detecting threats and by quite a significant margin (the maximum TPR obtained using PHEC is at least 10% more compared to any other model)[71]. Blockchain-based federated learning (BFL) is designed for privacy-awareness and efficient vehicular communication networking, where local on-vehicle machine learning (oVML) model updates are exchanged and verified in a decentralized way[42], [57]. Federated Deep Learning Framework for Privacy Preservation and Communication Efficiency FedPC, a Federated Deep Learning Framework for Communication Efficiency and Privacy Protection where CIFAR-10. LGG Segmentation dataset is used[61]. A Smart Factory's IoT-based system gives the hybrid model the ability to function effectively on deployed weak edge devices. The detecting work is divided up among smaller local zones in the final premises of traffic senders using the FL architectural design. As a result, anomalies or assaults may be swiftly found and contained in each zone. the researchers of [2] have used liquid storage data set FedeX-hybrid model based on VAE and SVDD FL-Based Explainable Anomaly Detection for ICS.

Table 4.2 is a collection of major attacks on CPS [58], [59], [60] and their summary, many attack types have the potential to seriously jeopardise the security and dependability of cyber-physical systems and federated learning. Poisoning attacks involve tampering with training data to distort machine learning outcomes, often difficult to detect. Communication attacks exploit flaws in system protocols, enabling data interception or manipulation. Inference attacks infer sensitive data from model outputs, posing privacy risks. Free-riding occurs when participants exploit federated learning without contributing, impacting system performance or data security. Defense strategies include data sanitization, encryption for secure communication, and robust optimization techniques[18], [32], [61], [62], [63]. Poisoning attacks include an attacker purposefully modifying or changing the training data used in machine learning models to provide inaccurate or misleading results. This form of assault can be used to impair essential system operations or to steal sensitive data. Poisoning attacks are especially difficult to identify and defend against because they might be difficult to differentiate from valid data. Communication attacks target flaws in the communication protocols used in cyberphysical systems and federated learning. These attacks can include eavesdropping, man-in-the-middle attacks, and other techniques that allow an attacker to intercept or manipulate the communication between different components of the system. Communication assaults can be used to steal sensitive data or impair system performance. Inference attacks include an attacker inferring sensitive information about the training data or the machine learning model by examining the model's output. This sort of attack can be used to steal sensitive data or to alter the model's behaviour. A free-riding attack occurs when a malevolent member in a federated learning system does not contribute their fair share of system resources (e.g., processing power, data) while still reaping the

Table 4.1: FL frameworks, their datasets, challenges and Performance metrics.

FL Architecture	Ref.	Model and Dataset	Description	Acc	Pre	FPR	TPR	Rec	F1-S	Challenges
Federated Averaging (FedAvg)	[45]	MNIST dataset	A communication-efficient approach for training deep neural networks in a decentralized manner.	0.9745	0.9695	0.0175	0.9745	0.9745	0.9720	Non-IID data distribution
FL-Based Explainable Anomaly Detection for ICS	[2]	VAE model with MNIST dataset	A framework for detecting anomalies in industrial control systems using Federated Learning and VAE models.	0.97	0.96	0.04	0.96	0.96	0.96	Privacy preservation, communication efficiency, explaining model decisions, dealing with imbalanced datasets and varying data distributions
Federated Deep Learning Framework for Privacy Preservation		Fashion-MNIST and CIFAR-10 datasets	A framework that preserves privacy by using a secure multi-party computation protocol in a decentralized environment.	-	-	-	-	-	-	Security, communication efficiency, privacy preservation, and scalability
Blockchain-based Federated Learning (BFL)	[61]	MNIST dataset	A framework that combines blockchain technology with FL to achieve security and privacy in a decentralized environment.	0.9896	0.9888	0.0112	0.9888	0.9888	0.9888	Security, communication efficiency, and privacy preservation
Federated Learning for Intrusion Detection in IoT Security	[69]	KDD Cup 99 dataset	A framework for intrusion detection in IoT security that uses ensemble learning and FL to improve detection accuracy.	0.9985	0.9868	0.0015	0.9868	0.9868	0.9868	Security, privacy preservation, and communication efficiency
Noise-Tolerant PHEC (NT-PHEC) in Federated Setup	[47]	MNIST dataset	A framework that uses NT-PHEC to deal with noisy labels and improve the accuracy of FL models.	0.9817	0.9739	0.0183	0.9739	0.9739	0.9739	Security, privacy preservation, communication efficiency, and dealing with noisy data
Federated Stochastic Gradient Descent (FedSGD)	[70]	Shakespeare dataset	A federated optimization algorithm for training machine learning models on decentralized data.	0.8598	0.8645	0.0235	0.8598	0.8598	0.8594	Network heterogeneity
Federated Averaging with Local Adaption (FedAvgLA)	[61]	CIFAR-10 dataset	An extension to FedAvg that adapts to local data by training a few extra local steps on each device.	0.8652	0.8675	0.0220	0.8652	0.8652	0.8652	Imbalanced data distribution
Federated Learning with Differential Privacy (FedDP)	[69]	EMNIST dataset	A framework for training deep learning models in a privacy-preserving manner by adding noise to gradients.	0.8996	0.8945	0.0190	0.8996	0.8996	0.8987	Privacy and utility trade-off
Secure Aggregation of Federated Learning (SAFL)	[71]	Facial recognition dataset	An approach that enables secure and privacy-preserving aggregation of model updates from multiple devices.	0.9772	0.9745	0.0105	0.9772	0.9772	0.9766	Communication and computation overheads
Federated Transfer Learning (FedTL)	[52], [72]	CUB-200 dataset	A federated learning framework for transferring knowledge from pre-trained models to similar but different tasks.	0.8256	0.8210	0.0338	0.8256	0.8256	0.8248	Task heterogeneity
Federated Multi-Task Learning (FedMTL)	[73]	Synthetic dataset	A federated approach for training models on multiple tasks in a decentralized setting.	0.9356	0.9335	0.0145	0.9356	0.9356	0.9347	Non-IID data and task heterogeneity
Federated Meta-Learning (FedMeta)	[74]	Omniglot dataset	A meta-learning approach for training models that can quickly adapt to new tasks in a federated setting.	0.9658	0.9625	0.0195	0.9658	0.9658	0.9652	Lack of labelled

Table 4.2: Types of major attacks, their source, and mitigation

Type of Attack	Name of Attack	Compromised	Source of Attack	Mitigation Techniques
Poisoning	Model Poisoning	Machine Learning Model	Data/Model Provider	Data Sanitization, Detection and Removal of Poisoned Data
	Data Poisoning	Training Data	Data Provider	Data Sanitization, Detection and Removal of Poisoned Data
	Gradient Manipulation	Machine Learning Model	Adversary	Robust Optimization Techniques
	Clean Label			Verification of Training Data and Model Outputs
	Dirty Label	Training Data	Data Provider	Data Sanitization, Detection and Removal of Poisoned Data
	Training Rule Manipulation	Machine Learning Model	Adversary	Detection of Anomalous Model Behaviour, Use of Secure and Trusted Algorithms
	Backdoor			Regular Monitoring of Model Behaviour, Robust Optimization Techniques
Communication	MITM	Communication Channel		Secure Communication Protocols, Encryption
	Communication Bottlenecks		Network Infrastructure	Network Optimization Techniques
	Evasion Attacks	Machine Learning Model	Adversary	Use of Adversarial Training, Detection and Removal of Adversarial Examples
Inference	Membership Inference	Machine Learning Model		Use of Differential Privacy, Randomized Response
	Properties Inference			Verification of Model Outputs
	Training Inputs Inference			Use of Differential Privacy, Randomized Response
	Label Inference			Verification of Training Data and Model Outputs
	GANs based Inference			Use of Adversarial Training, Detection and Removal of Adversarial Examples
Free Riding	Data Free Riding	Data Provider	Participant	Secure Aggregation, Incentives and Penalties for Participants
	Model Free Riding	Model Provider		Secure Aggregation, Incentives and Penalties for Participants

advantages of the trained model. This form of attack can be used to either impair system operation or steal sensitive data Poisoning attacks include an attacker purposefully modifying or changing the training data used in machine learning models to provide inaccurate or misleading results. This form of assault can be used to impair essential system operations or to steal sensitive data. Poisoning attacks are especially difficult to identify and defend against because they might be difficult to differentiate from valid data. Communication attacks target flaws in the communication protocols used in cyberphysical systems and federated learning. These attacks can include eavesdropping, man-in-the-middle attacks, and other techniques that allow an attacker to intercept or manipulate the communication between different components of the system. Communication assaults can be used to steal sensitive data or impair system performance. Inference attacks include an attacker inferring sensitive information about the training data or the machine learning model by examining the model's output. Training Data and Model Outputs GANs based Inference Use of Adversarial Training, Detection and Removal of Adversarial Examples Free Riding Data Free Riding Data Provider Participant Secure Aggregation, Incentives and Penalties for Participants Model Free Riding Model Provider Secure Aggregation, Incentives and Penalties for Participants necessitates a thorough understanding of these threats as well as the strategies employed by attackers to exploit weaknesses in cyber-physical systems and federated learning.

5. Results and discussions. The study emphasizes how crucial it is to secure Cyber-Physical Systems (CPS). The need to protect CPS against cyber threats has grown as a result of its widespread use in industries like power grids, transportation networks, and healthcare institutions. When it comes to CPS security, traditional machine learning techniques encounter many obstacles, including as the requirement to centralize data for model training, scalability problems, and privacy issues. Federated learning (FL) shows promise as a way to address these issues. FL reduces the dangers related to centralized data processing and storage by facilitating cooperative model training over decentralized edge devices while protecting data privacy. The paper shows how FL approaches optimize model performance across many contexts by dividing up the learning process among several edge devices.

6. Conclusion. In conclusion, the study highlights the critical importance of securing cyber-physical systems (CPS) in today's interconnected world. As we navigate an era where digital threats loom large over infrastructure, our study underscores the significance of addressing vulnerabilities and ensuring the integrity of data exchanges within CPS environments. Through an exploration of federated learning (FL) architectures, the study has presented a viable solution to enhance the security and privacy of these systems. By embracing FL models, we mitigate concerns surrounding centralized data storage and processing, thus reducing the risk of single-point failures. Our analysis demonstrates FL's potential in bolstering security protocols while safeguarding sensitive information across distributed entities. Furthermore, our investigation into FL's application in intrusion detection within CPS underscores its capacity to proactively mitigate emerging threats, including zero-day attacks. Looking ahead, future research efforts should concentrate on refining FL methodologies tailored specifically for CPS security. This involves extensive training with diverse datasets and real-world scenarios to fortify FL models' efficacy in detecting and mitigating threats. Additionally, the development of adaptive intrusion detection systems capable of swift response to evolving attack vectors will be paramount. By advancing FL techniques and seamlessly integrating them into CPS security frameworks, we pave the way for a more resilient infrastructure, ensuring the safeguarding of economic assets, human lives, and the integrity of our critical systems. In essence, our findings emphasize the transformative potential of federated learning in fortifying CPS security, setting a precedent for continued innovation and collaboration in safeguarding our digital future.

REFERENCES

[1] E. Monmasson and M. Cirstea, "FPGA Design Methodology for Industrial Control Systems – a Review".

[2] T. T. Huong et al., "Federated Learning-Based Explainable Anomaly Detection for Industrial Control Systems," IEEE Access, vol. 10, pp. 53854–53872, 2022, doi: 10.1109/ACCESS.2022.3173288.

[3] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-Physical Systems Security - A Survey," IEEE Internet Things J., vol. 4, no. 6, pp. 1802–1831, 2017, doi: 10.1109/JIOT.2017.2703172.

[4] E. A. Lee, "Cyber physical systems: Design challenges," Proc. - 11th IEEE Symp. Object/Component/Service-Oriented Real-Time Distrib. Comput. ISORC 2008, no. August, pp. 363–369, 2008, doi: 10.1109/ISORC.2008.25.

[5] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Commun. Surv. Tutorials, vol. 17, no. 4, pp. 2347–2376, 2015, doi: 10.1109/COMST.2015.2444095.

[6] R. Alguliyev, Y. Imamverdiyev, and L. Sukhostat, "Cyber-physical systems and their security issues," Comput. Ind., vol. 100, no. July 2017, pp. 212–223, 2018, doi: 10.1016/j.compind.2018.04.017.

[7] M. K. Meng, "An innovative industrial control system architecture for real-time response , fault-tolerant operation and seamless plant integration," no. June, pp. 569–581, 2021, doi: 10.1049/tje2.12064.

[8] H. D. Gómez, J. Garcia-Rodriguez, J. Azorin-Lopez, D. Tomás, A. Fuster-Guillo, and H. Mora-Mora, "IA-CPS: Intelligent architecture for cyber-physical systems management," J. Comput. Sci., vol. 53, no. April, p. 101409, 2021, doi: 10.1016/j.jocs.2021.101409.

[9] M. Wolf and D. Serpanos, "Safety and security in cyber-physical systems and internet-of-things systems," Proc. IEEE, vol. 106, no. 1, pp. 9–20, 2018, doi: 10.1109/JPROC.2017.2781198.

[10] Y. Zacchia Lun, A. D'Innocenzo, F. Smarra, I. Malavolta, and M. D. Di Benedetto, "State of the art of cyber-physical systems security: An automatic control perspective," J. Syst. Softw., vol. 149, pp. 174–216, 2019, doi: 10.1016/j.jss.2018.12.006.

[11] J. Lee, B. Bagheri, and H. Kao, "ScienceDirect A Cyber-Physical Systems architecture for Industry 4 . 0-based manufacturing systems," Manuf. Lett., vol. 3, pp. 18–23, 2015, doi: 10.1016/j.mfglet.2014.12.001.

[12] J. Qian, X. Du, B. Chen, B. Qu, K. Zeng, and J. Liu, "Cyber-Physical Integrated Intrusion Detection Scheme in

SCADA System of Process Manufacturing Industry," IEEE Access, vol. 8, pp. 147471–147481, 2020, doi: 10.1109/ACCESS.2020.3015900.

[13] P. Cassara, A. Gotta, and L. Valerio, "Federated Feature Selection for Cyber-Physical Systems of Systems," IEEE Trans. Veh. Technol., vol. 71, no. 9, pp. 9937–9950, 2022, doi: 10.1109/tvt.2022.3178612.

[14] C. C. Sun, C. C. Liu, and J. Xie, "Cyber-physical system security of a power grid: State-of-the-art," Electron., vol. 5, no. 3, 2016, doi: 10.3390/electronics5030040.

[15] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams, and A. Hahn, "Guide to Industrial Control Systems (ICS) Security NIST Special Publication 800-82 Revision 2," NIST Spec. Publ. 800-82 rev 2, pp. 1–157, 2015, [Online]. Available: http://industryconsulting.org/pdfFiles/NIST Draft-SP800-82.pdf

[16] M. Conti, S. Member, I. D. Donadel, and F. Turrin, "A Survey on Industrial Control System Testbeds and Datasets for Security Research," 2017.

[17] L. Rosa, M. Freitas, and S. Mazo, "A Comprehensive Security Analysis of a SCADA Protocol: From OSINT to Mitigation," vol. 7, 2019, doi: 10.1109/ACCESS.2019.2906926.

[18] F. O. Olowononi, D. B. Rawat, and C. Liu, "Resilient Machine Learning for Networked Cyber Physical Systems: A Survey for Machine Learning Security to Securing Machine Learning for CPS," IEEE Commun. Surv. Tutorials, vol. 23, no. 1, pp. 524–552, 2021, doi: 10.1109/COMST.2020.3036778.

[19] L. Cao, X. Jiang, Y. Zhao, S. Wang, D. A. N. You, and X. Xu, "A Survey of Network Attacks on Cyber-Physical Systems," pp. 44219–44227, 2020.

[20] E. Irmak, "An overview of cyber-attack vectors on SCADA systems," no. August, 2022, doi: 10.1109/ISDFS.2018.8355379.

[21] N. Y. Kim, S. Rathore, J. H. Ryu, J. H. Park, and J. H. Park, "A Survey on Cyber Physical System Security for IoT: Issues , Challenges , Threats , Solutions," vol. 14, no. 6, pp. 1361–1384, 2018.

[22] A. Daneels and W. Salter, "What Is Scada?," Int. Conf. Accel. Large Exp. Phys. Control Syst. Trieste, Italy, pp. 339–343, 1999, [Online]. Available: http://scholar.google.com/scholar?hl=enbtnG=Searchq=intitle:WHAT+IS+SCADA+?0

[23] D. Pliatsios, P. Sarigiannidis, T. Lagkas, and A. G. Sarigiannidis, "A Survey on SCADA Systems: Secure Protocols, Incidents, Threats and Tactics," IEEE Commun. Surv. Tutorials, vol. 22, no. 3, pp. 1942–1976, 2020, doi: 10.1109/COMST.2020.2987688.

[24] S. Samtani, S. Yu, H. Zhu, M. Patton, and H. Chen, "Identifying SCADA Vulnerabilities Using Passive and Active Vulnerability Assessment Techniques," no. September, 2016, doi: 10.1109/ISI.2016.7745438.

[25] B. A. Salau, A. Rawal, and D. B. Rawat, "Recent Advances in Artificial Intelligence for Wireless Internet of Things and Cyber-Physical Systems: A Comprehensive Survey," IEEE Internet Things J., vol. 9, no. 15, pp. 12916–12930, 2022, doi: 10.1109/JIOT.2022.3170449.

[26] L. Fumagalli, E. Negri, O. Severa, P. Balda, and E. Rondi, "Distributed control via modularized CPS architecture Lessons learnt from an industrial case study," IFAC-PapersOnLine, vol. 51, no. 11, pp. 803–808, 2018, doi: 10.1016/j.ifacol.2018.08.417.

[27] F. Akbarian, W. Tarneberg, E. Fitzgerald, and M. Kihl, "Attack Resilient Cloud-Based Control Systems for Industry 4.0," IEEE Access, vol. 11, no. March, pp. 27865–27882, 2023, doi: 10.1109/ACCESS.2023.3259063.

[28] D. Gonzalez, F. Alhenaki, and M. Mirakhorli, "Architectural security weaknesses in industrial control systems (ICS) an empirical study based on disclosed software vulnerabilities," Proc. - 2019 IEEE Int. Conf. Softw. Archit. ICSA 2019, pp. 31–40, 2019, doi: 10.1109/ICSA.2019.00012.

[29] H. Zhu, H. Zhang, and Y. Jin, "From federated learning to federated neural architecture search: a survey," Complex Intell. Syst., vol. 7, no. 2, pp. 639–657, 2021, doi: 10.1007/s40747-020-00247-z.

[30] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Access, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 140699–140725, 2020. doi: 10.1109/ACCESS.2020.3013541.

[31] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.04885

[32] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," IEEE Communications Surveys and Tutorials, vol. 23, no. 3. Institute of Electrical and Electronics Engineers Inc., pp. 1622–1658, Jul. 01, 2021. doi: 10.1109/COMST.2021.3075439.

[33] K. Wei et al., "Vertical Federated Learning: Challenges, Methodologies and Experiments," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.04309

[34] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.04977

[35] Q. Wu, K. He, and X. Chen, "Personalized Federated Learning for Intelligent IoT Applications: A Cloud-Edge based Framework," IEEE Comput. Graph. Appl., pp. 1–9, 2020, doi: 10.1109/OJCS.2020.2993259.

[36] C. Wang, G. Yang, G. Papanastasiou, H. Zhang, J. J. P. C. Rodrigues, and V. H. C. De Albuquerque, "Industrial Cyber-Physical Systems-Based Cloud IoT Edge for Federated Heterogeneous Distillation," IEEE Trans. Ind. Informatics, vol. 17, no. 8, pp. 5511–5521, 2021, doi: 10.1109/TII.2020.3007407.

[37] S. K. Lo, Q. Lu, L. Zhu, H. Y. Paik, X. Xu, and C. Wang, "Architectural patterns for the design of federated learning systems," J. Syst. Softw., vol. 191, pp. 1–19, 2022, doi: 10.1016/j.jss.2022.111357.

[38] H. Ludwig, Federated Learning. 2022. doi: 10.1007/978-3-030-96896-0.

[39] S. Feng and H. Yu, "Multi-Participant Multi-Class Vertical Federated Learning," 2020, [Online]. Available: http://arxiv.org/abs/2001.11154

[40] T. D. Cao, T. Truong-Huu, H. Tran, and K. Tran, "A federated learning framework for privacy-preserving and parallel training," arXiv, no. January, 2020.

[41] D. Liu, T. Miller, R. Sayeed, and K. D. Mandl, "FADL:Federated-Autonomous Deep Learning for Distributed Electronic Health Record," 2018, [Online]. Available: http://arxiv.org/abs/1811.11400

[42] S. R. Pokhrel and J. Choi, "Federated Learning with Blockchain for Autonomous Vehicles: Analysis and Design Challenges," IEEE Trans. Commun., vol. 68, no. 8, pp. 4734–4746, Aug. 2020, doi: 10.1109/TCOMM.2020.2990686.

[43] P. Consul, I. Budhiraja, R. Chaudhary, and D. Garg, "FLBCPS: Federated Learning based Secured Computation Offloading in Blockchain-Assisted Cyber-Physical Systems," Proc. - 2022 IEEE/ACM 15th Int. Conf. Util. Cloud Comput. UCC 2022, pp. 412–417, 2022, doi: 10.1109/UCC56403.2022.00071.

[44] L. T. Yang, R. Zhao, D. Liu, W. Lu, and X. Deng, "Tensor-Empowered Federated Learning for Cyber-Physical-Social Computing and Communication Systems," IEEE Commun. Surv. Tutorials, vol. 25, no. 3, pp. 1909–1940, 2023, doi: 10.1109/COMST.2023.3282264.

[45] Z. Lian et al., "DEEP-FEL: Decentralized, Efficient and Privacy-Enhanced Federated Edge Learning for Healthcare Cyber Physical Systems," IEEE Trans. Netw. Sci. Eng., vol. 9, no. 5, pp. 3558–3569, 2022, doi: 10.1109/TNSE.2022.3175945.

[46] J. Cui et al., "Collaborative Intrusion Detection System for SDVN: A Fairness Federated Deep Learning Approach," IEEE Trans. Parallel Distrib. Syst., vol. 34, no. 9, pp. 2512–2528, 2023, doi: 10.1109/TPDS.2023.3290650.

[47] M. Chahoud et al., "ON-DEMAND-FL: A Dynamic and Efficient Multi-Criteria Federated Learning Client Deployment Scheme," IEEE Internet Things J., vol. 10, no. 18, pp. 15822–15834, 2023, doi: 10.1109/JIOT.2023.3265564.

[48] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems," IEEE Trans. Ind. Informatics, vol. 17, no. 8, pp. 5615–5624, 2021, doi: 10.1109/TII.2020.3023430.

[49] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber–Physical Systems," IEEE Trans. Ind. Informatics, vol. 17, no. 8, pp. 5615–5624, Aug. 2021, doi: 10.1109/TII.2020.3023430.

[50] K. Demertzis, "Blockchained Federated Learning for Threat Defense," pp. 1–12, 2021, [Online]. Available: http://arxiv.org/abs/2102.12746

[51] S. H. Javed et al., "APT Adversarial Defence Mechanism for Industrial IoT Enabled Cyber-Physical System," IEEE Access, vol. 11, no. June, pp. 74000–74020, 2023, doi: 10.1109/ACCESS.2023.3291599.

[52] S. You, "A Cyber-secure Framework for Power Grids Based on Federated Learning," pp. 1–4.

[53] T. D. Cao, T. Truong-Huu, H. Tran, and K. Tran, "A federated deep learning framework for privacy preservation and communication efficiency," J. Syst. Archit., vol. 124, 2022, doi: 10.1016/j.sysarc.2022.102413.

[54] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning," IEEE Access, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.

[55] A. Zainudin, R. Akter, D. S. Kim, and J. M. Lee, "Federated Learning Inspired Low-Complexity Intrusion Detection and Classification Technique for SDN-Based Industrial CPS," IEEE Trans. Netw. Serv. Manag., vol. PP, p. 1, 2023, doi: 10.1109/TNSM.2023.3299606.

[56] S. Chatterjee and M. K. Hanawal, "Federated Learning for Intrusion Detection in IoT Security: A Hybrid Ensemble Approach".

[57] T. Zhang, C. He, T. Ma, L. Gao, M. Ma, and S. Avestimehr, "Federated Learning for Internet of Things," SenSys 2021 - Proc. 2021 19th ACM Conf. Embed. Networked Sens. Syst., vol. 23, no. 3, pp. 413–419, 2021, doi: 10.1145/3485730.3493444.

[58] V. Casola, A. De Benedictis, C. Mazzocca, and R. Montanari, "Designing Secure and Resilient Cyber-Physical Systems: a Model-based Moving Target Defense Approach," IEEE Trans. Emerg. Top. Comput., vol. PP, no. X, pp. 1–12, 2022, doi: 10.1109/TETC.2022.3197464.

[59] L. M. Castiglione and E. C. Lupu, "Which Attacks Lead to Hazards? Combining Safety and Security Analysis for Cyber-Physical Systems," IEEE Trans. Dependable Secur. Comput., vol. PP, pp. 1–16, 2023, doi: 10.1109/TDSC.2023.3309778.

[60] Y. Song, T. Liu, T. Wei, X. Wang, Z. Tao, and M. Chen, "FDA3: Federated Defense against Adversarial Attacks for Cloud-Based IIoT Applications," IEEE Trans. Ind. Informatics, vol. 17, no. 11, pp. 7830–7838, 2021, doi: 10.1109/TII.2020.3005969.

[61] G. K. Pandey, D. S. Gurjar, H. H. Nguyen, and S. Yadav, "Security Threats and Mitigation Techniques in UAV Communications: A Comprehensive Survey," IEEE Access, vol. 10, no. October, pp. 112858–112897, 2022, doi: 10.1109/ACCESS.2022.3215975.

[62] P. Asef, R. Taheri, M. Shojafar, I. Mporas, and R. Tafazolli, "SIEMS: A Secure Intelligent Energy Management System for Industrial IoT Applications," IEEE Trans. Ind. Informatics, vol. 19, no. 1, pp. 1039–1050, 2023, doi: 10.1109/TII.2022.3165890.

[63] M. Benmalek, M. A. Benrekia, and Y. Challal, "Security of Federated Learning: Attacks, Defensive Mechanisms, and Challenges," Rev. d'Intelligence Artif., vol. 36, no. 1, pp. 49–59, 2022, doi: 10.18280/RIA.360106.

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Mar 6, 2024
Accepted: May 29, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 34–47, DOI 10.12694/scpe.v26i1.3479

A GPS-ENABLED FUEL SENSOR BASED VEHICLE TRACKING SYSTEM FOR FLEET MANAGEMENT USING THE INTERNET OF THINGS

A. SRINAGESH,[*] CH. APARNA ,[†] AND M.V.P. CHANDRA SEKHARA RAO[‡]

Abstract. The key objective of this paper is to monitor the Fuel consumption, route deviations, driving habits, breakdown details, vehicle maintenance conditions, and other vital performance details of a Heavy vehicle especially a college bus using a pre-fitted fuel sensor needed for an extensive fleet management system as operational costs and maintenance are escalating day by day. In particular, vehicle owners are confronting fuel break-ins, spare parts theft and illegal parking or route diversions in transport vehicles. In addition to this, vehicle proprietors don't ascertain operational details daily if they are illiterate. Various models are available which Standalone, Onboard diagnostics-based, GPS GPS-based fuel tracking Systems. This research project leverages IoT technology, GPS-based sensors, and a GSM tracking system to provide vehicle owners with real-time information on fuel consumption. This cloud-based and mobile application tracks the vehicle in real time. Most of the models are web-based and custom-built only. When the driver initiates fuel filling, the Fuel sensor in the tank automatically activates and senses the amount of fuel in the diesel tank by transmitting fuel input data to a connected cloud web server for storage and analysis along with other data. The gathered values are cross-verified with the database, and an alert message is promptly sent to the vehicle owner based on consumption patterns. This mobile app-based Vehicle Tracking System (VTS) application, using SMS and Vehicle tracking features offers numerous advantages to the vehicle owner, primarily focusing on preventing fuel theft.

Key words: Fuel Sensor, GPS Sensor, Ultra Sound Sensor, Arduino Uno, ESP 8266, Google Maps, Internet of Things

1. Introduction. The Internet of Things (IoT) network is composed of computing equipment, actuating mechanical and digital devices, and interconnected sensors. These devices, along with objects or individuals equipped with unique identifiers, can exchange real-time data over a network autonomously without requiring direct human intervention. Specifically, fuel management systems are designed to accurately measure and control fuel consumption in various sectors, including transportation and construction.

The Internet of Things (IoT) facilitates the regulation and communication with computer systems. In the era of IoT, Vehicle Tracking Systems (VTS) have gained significant importance. VTS applications are diverse, from route tracking and vehicle monitoring to component maintenance. VTS offers opportunities to integrate various new technologies for an enhanced On-Board Diagnostics (OBD) experience, particularly in virtual environments. OBD and GPS-enabled navigation systems have been actively researched in recent years. Within the realm of VTS systems, Sensor-based VTS stands out as the most organic, user-friendly, and intuitive means of facilitating communication between humans and machines, mirroring the patterns of human interaction. Its intuitive nature has resulted in widespread applications for navigating extensive and intricate data in fleet management. Fuel sensing and tracking have long been intriguing challenges in the Vehicle Tracking Systems community. This is primarily due to the significant fuel costs of transporting goods from one location to another. Ensuring minimal maintenance and vehicle downtime is crucial for fleet owners, presenting a real-time challenge. The main hurdle lies in the semantic gap between visually inspecting a vehicle and receiving real-time data from embedded sensors. While humans can often detect issues through experience or intuition, computers rely solely on live data.

Vehicle Tracking System (VTS) with automatic fuel Detection is meant to acknowledge various fuel consumption details and has become a valuable and mandatory component in Vehicles. Fuel sensors are essential components of automobile dashboards that provide straightforward and standard information. GPS-enabled Fuel Sensors have received much research attention in recent years. However, the sector still presents a variety

[*] Department of CSE, R.V.R & J.C. College of Engineering, Guntur, India (asrinagesh@gmail.com)

[†] Department of CSE, R.V.R & J.C. College of Engineering, Guntur, India (achaparala@gmail.com)

[‡] Department of CSE, R.V.R & J.C. College of Engineering, Guntur, India (manukondach@gmail.com)

of challenges for researchers. Fuel tracking facilitated by GPS processes additional data. The accumulation of data and interpretation of the route must be executed with precision and timeliness. The proposed is a low-cost VTS system with fuel tracing. Location detection and Google Maps API key-based tracking detection capture accurate locations from the background regions. A connected ultra-sound-based fuel sensing is also suggested to compare the fuel consumption in the vehicle. The primary objective of this system is to analyze fuel consumption by comparing the amount of gasoline used on a certain route with the data collected from a sensor. The system then converts this information into text alarm messages, which are sent to the owner of the vehicle. The Mobile Application provides an easy method of tracking fuel consumption.

The study aims to track the route and the consumption information of a fleet management system and detect fuel consumption utilizing a pre-fitted fuel sensor to the vehicle. People are dealing with gasoline theft in transport vehicles due to fuel price increases. Additionally, if a car owner is illiterate, they cannot determine how much fuel their vehicle needs daily. By employing this system, the owner of a car will be aware of the fuel use. For example, when the driver starts to fill up the tank with fuel, the ultrasonic and level sensors activate and record data for the portable application. A low-cost VTS system with Fuel tracking is proposed. Location detection and Google Maps API key-based tracking detection capture accurate locations from the background regions. A connected ultrasound-based fuel sensing is also suggested to compare the fuel consumption in the vehicle. The main aim of this system is to compare the fuel consumed based on the route travelled with the fuel consumed from a sensor and convert them into corresponding text alert messages to the vehicle's owner. The Mobile Application provides an easy method of tracking fuel consumption.

Vehicle Tracking Systems (VTS) can encompass a single function or a series of tasks within standard vehicle management, either static or dynamic. VTS systems must deliver accurate data promptly. Implementing a fuel detection system using GPS poses a significant challenge, as fuel sensors must be carefully deployed due to the high flammability of fuel. This complexity arises from various factors, including the diverse environmental conditions, vehicle status, flexible system designs, and the need for real-time execution. It is possible to save the information that is received by sensors in computerized systems , which enables the development of reports that include useful information to implement informed fuel management practices. This facilitates control over utilization, cost analysis, and accurate expense tracking related to fuel purchases. While contemporary vehicle tracking systems often utilize GPS technology for location tracking, other advancements in automatic vehicle location within the Internet of Things (IoT) can also be employed. The information collected from vehicles is crucial for identifying static and dynamic data changes and anticipating driver behaviors' and capabilities shifts.

There are various approaches to acquiring a Vehicle Information Securing Framework, including a compact system relying on standard hardware components such as a computer or CAN/USB interface for sensed data. The framework involves critical components like:

- KI: A database between the instrument panel and the vehicle Gateway.
- INFOTAINMENT: A data bus connecting entertainment devices (radio, amplifier, CD changer, etc.).
- ENGINE: A data bus connecting different engine parts (drive, ABS, transmission, wheel, etc.).
- COMFORT: A data bus linking comfort devices (air conditioner, door controller, central control unit, navigation panel, etc.).
- KLINE: A diagnostic bus.

Collectively, these components contribute to a comprehensive system for effectively acquiring and managing vehicle information.

2. Literature review. Chiwhane, S.A., et al. introduced in [1] a system to prevent fraud at petrol pumps. Their system involves a flow sensor that activates when fueling begins, providing pulses proportional to the flow rate, which are sent to a cloud server via ESP8266. The user's location is also tracked using GPS via a user application. Unlike earlier approaches that utilized a flow sensor, the proposed system utilizes an ultrasonic sensor to measure fuel levels in various dimensions of fuel tanks.

Padmaja, B.V. et al. developed in [2] a vehicle tracking system using Blynk for data transport and visualization . The system has Ultrasonic, Gas, IR, Temperature, and GPS sensors. The suggested system monitors via mobile app.

Dukare, S.S., Patil, D.A., and Rane, K.P. introduced in [3] a system for vehicle tracking, monitoring, and alerting. The alerting system transmits information via GSM or GPRS, while GPS provides the vehicle's specific position. In contrast, the suggested system goes beyond these functions by including gasoline monitoring, vehicle position tracking, identifying the closest fuel stations, and getting alarm alerts.

Gullipalli, Karri, and Kota introduced in [4] a system that incorporates an Arduino, GPS, GSM, a fuel sensor, and a speed sensor This system facilitates communication and data exchange between the devices on the bus, web applications, and desktop applications. The researchers utilized NodeMCU (ESP8266) as a key component of their proposed system. It also uses a web app to monitor the system, whereas the suggested solution uses a mobile app.

Ribeiro and Gonzaga presented in [5] several approaches for real-time background removal algorithms based on the Gaussian Mixture Method (GMM) employing video sequences for image segmentation.

Vanmore, S.V. et al. developed in [6] a GPS/GSM vehicle tracking and location system . GPS tracking reports allow this system to follow the vehicle's status. Android apps provide vehicle tracking for safety.

Alshamisi and K'epuska proposed in [7] a vehicle tracking system that utilizes GPS and GSM technologies. Their solution employs a GPS and a GSM modem, which are connected to a vehicle by an Arduino MEGA2560. On the other hand, the suggested solution chooses to use NodeMCU (ESP8266) which has an integrated Wi-Fi chip for transmitting data.

Rohitaksha, K., Madhu, C.G., Nalini, B.G., and Nirupama, C.V. created in [8] a parallel processor system. The user interface on another Android phone lets people follow a vehicle on Google Maps. The suggested system includes fuel monitoring and vehicle activity tracking, whereas the current system just tracks vehicle position.

Khin and Oo. recommended in [9] utilizing Arduino, GPS, GSM, and web-based technologies . The proposed method uses an online fireplace database, whereas the realized system uses an online MySQL database server. The evaluated study utilizes a web application for system monitoring, whereas the suggested method uses a mobile app.

Saini, J., Agarwal, M., Gupta, A., and Manjula, R. introduced in [10] a car tracking system that utilizes GPS and GSM technology via an Android app. The main objective of the system is to monitor the precise position of the vehicle. The suggested system utilizes NodeMCU to establish communication and transmit data to the user.

In the current research, most of the automatic fuel detection models using Mobile Application and Cloud-based data storage with Firebase technology are implemented in this paper that is custom-built or tailor-made to suit with useful and key functional features for a vehicle. The main gap is that the Software application can be designed, developed, and deployed in different technologies based on the requirements of the user. These software applications are web-based, Cloud-based, IoT enabled only. In this domain, advanced IoT and cloud-based models with more numbers are still in the nascent stage for moderate fleet management scenarios.

3. Proposed Methodology.

Data Authentication. The first module in the proposed system is the Registration module. A driver has to register with his Mobile number as UNIQUEID. Once registered it will authenticate and the tracking application presents a user interface to track basic vehicle data only.

Data Input and Pre-processing. Some basic input details like the purpose of travel. Source and destination along with the Vehicle Number, Date of Fitness Certificate, and some mandatory fields have to be entered the First time and once only.

Data Captured from the Sensors. Mainly a GPS-enabled Fuel Sensor is fitted in the vehicle and complete Fuel and Route information is tracked from that instance continuously. The data consumed for this purpose is also displayed.

System architecture. The system architecture, as illustrated in Figure 3.1, functions as a conceptual model that defines the system's structure, behavior, and perspectives. To fabricate this model, an ultrasonic sensor is utilized initially to measure the fuel level, after which the value is converted to volume. The NodeMCU (ESP8266) module is utilized to transmit this data to the server, whereas the GPS module provides the latitude and longitude of the vehicle. By aggregating all incoming data and displaying it through the mobile application interface, the server implements a system for location tracking and real-time monitoring.

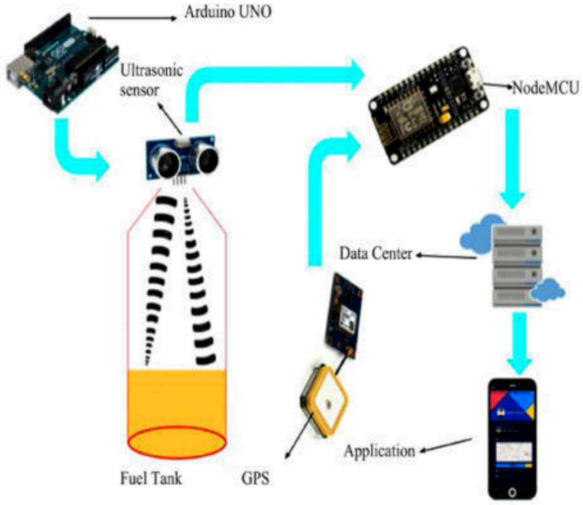


Fig. 3.1: Architecture of the Proposed System

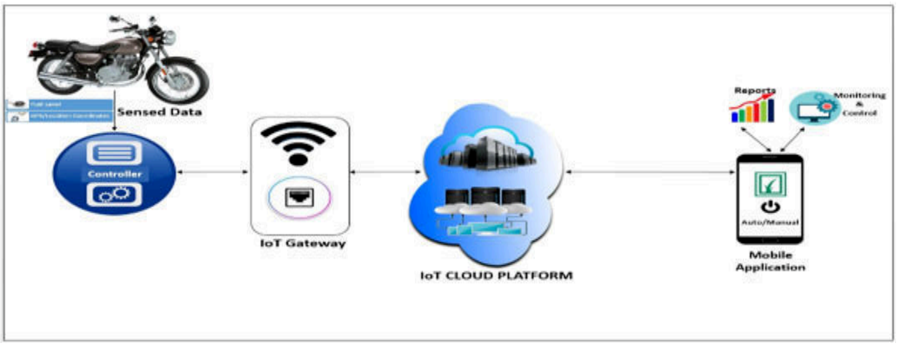


Fig. 4.1: Block diagram of the proposed method

4. Framework Design. A Framework design is a conceptual model that outlines a system's structure, behavior, and aspects, as depicted in Figure . To build this model, the first step is using an ultrasonic sensor to gauge the fuel level and then turning it into volume. Afterward, the data is sent to the server via the NodeMCU (ESP8266) module, while the GPS module transmits the latitude and longitude of the vehicle to the server. The server gathers and displays all of this data in the mobile application, enabling real-time monitoring via a global positioning system.

- Vehicles go outside the path of travel due to various reasons that are not intended or directed or related locations to driver behaviour and other personal reasons and arrive late to the destination. Do not report to a destination location within the set time indicated. That's why it was necessary to carry out a GPS Mapping.
- Different sensors were integrated with the vehicle that was to be tracked. Visually identical, they were only differentiated by being located in the wrong place. They needed to be identified.
- Fuel or vehicle conditions caused the reduction of the overutilization of the Fuel. Because of that, such cost was confused with the operational or maintenance of the vehicle, which was younger than the vehicle's life.
- It was necessary to include activities to the equivalent damage due to abnormal traffic and extreme

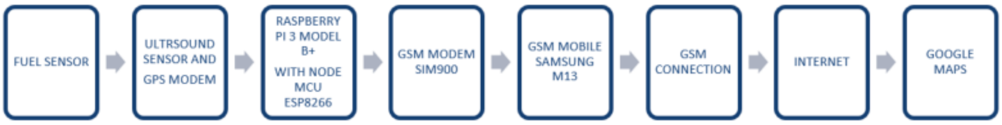


Fig. 4.2: Flow diagram of the proposed method

Fuel Management Section	**Vehicle Management System Components**
Raspberry Pi	Fuel Level Sensor (2)
NodeMCU (ESP8266)	Vehicle Speedometer Sensor
Ultrasound Sensor	GPS Module
PIR Sensor	Wi-Fi Module
Buzzer	Wide-angle camera

Table 4.1: List of Components

Fig. 6.1: Raspberry Pi 3 Model B+

road conditions corresponding to the route within the range of travel. The issue was addressed using Google Maps, incorporating authorized waybills provided by the client.

5. Fuel and GPS Sensors.

Fuel Sensor. The fuel tank pressure sensor, integral to the fuel pump assembly, is typically situated on top of or inside the fuel tank. This component is crucial to the evaporative emissions system (”EVAP”). Its primary function is to measure pressure within the fuel system, enabling the detection of evaporative leaks, such as those caused by a loose or malfunctioning gas cap. A concrete instance is the Lawrence Fuel Flow Sensor, which includes a 10-foot cable and a T-connector.

GPS Sensor. This global positioning system (GPS) is a satellite-based navigation system that utilizes a network of 24 satellites circling the Earth. GPS sensors, which are outfitted with antennas, function as receivers for this system. These sensors provide accurate information on position, velocity, and timing. A concrete example is the Globalstar 9600 Satellite Data Hotspot data.

6. Software and Hardware Specifications.

6.1. Hardware Requirements. The Raspberry Pi 3 Model B+, which is shown in Figure 6.1, is one of the physical components of the system. Other components include a Camera Board, a 5-inch 800x480 Resistive HD Touch Screen, an L298 H-bridge driver, a four-wheel-drive Rover chassis, rechargeable batteries for power delivery, and an ultrasonic sensor.

6.2. Software Platform. The Vehicle Tracking System is implemented using Python on the Intel Core TM i5 Processor. The Raspbian operating system is required. The required programs for the system may be summarized as follows:

Raspbian Operating System. The Raspbian OS is specifically designed for the Raspberry Pi. Its repository, comprising over 35,000 packages, provides comprehensive support for the Raspberry Pi environment. This operating system is freely available for download from the internet, commonly referred to as NOOBS, and can be subsequently transferred onto a 16GB (or larger) RAM stick.

Python and Libraries. Python is a popular general-purpose programming language. First developed in 1991. It lets programmers express ideas in fewer lines than C or Java. Python is a dynamic system that allows object-oriented, functional, and procedural programming with automated memory management. Python code can execute on many operating systems owing to its extensive and comprehensive standard libraries and Python interpreters.

Assumptions Made. To perform the Software and hardware modules what are the main assumptions have been considered needed to represent at the start of the results and discussion. The Fuel and Ultrasound Sensors are properly connected, interfaces with GSM and GPS are established, and all hardware configuration settings are set up. Additionally, it is assumed that the required Python libraries are available to implement this experiment.

7. Fuel Sensor Sensing with Level Measurement Implementation and Applications. The steps are the followings:

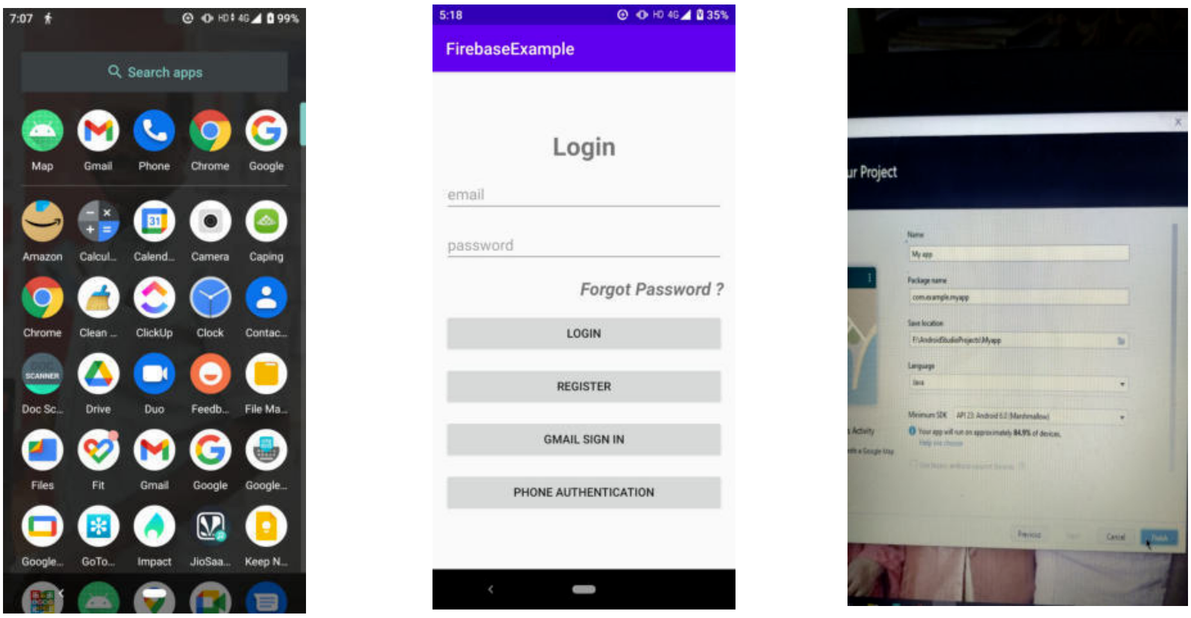
1. Import necessary GPS and GSM sensor packages in Raspberry Pi, for example, import RPi.GPIO as GPIO package.
2. Define the GPIO pins of the Raspberry Pi to connect it to a mobile.
3. Initialize the variable.
4. Capturing is done as follows: For distance inside the Fuel tank.
5. Compute the Fuel Consumption, Distance travelled, and Balance Fuel in the vehicle.
6. If any fuel is misused, stolen, or lost, send an SMS or the Fuel info message daily.
7. Find any route deviation using Google Maps.
8. Generate Reports.
   For Example: Find distance using an ultrasonic sensor as follows:

```
dist = round(dist, 2)
distance = avgdistance + dist
if avgdistance < 15:
    stop()
    backward()
else:
    forward()
```

8. Experimental Results and Discussion. To begin the process of managing the four DC motors that are installed on the mobile robot, the first step is to connect each motor to the A (Out 1 & Out 2) and B (Out 3 & Out 4) connectors that are located on the L298N module. The L298N module is then powered by two 9V batteries once this step has been completed. In the meanwhile, the Raspberry Pi requires a 5V intelligent supply to function properly.

The Raspberry Pi utilizes six GPIO pins for motor control. GPIO10 controls motor A, GPIO09 controls motor B, and the input pins (IN1, IN2, IN3, and IN4) of the L298N driver are linked to GPIO22, GPIO18, GPIO16, and GPIO12 of the Raspberry Pi, respectively. The Ultrasonic sensor consists of four pins: VCC, which is linked to GPIO 5V (pin 2); GND, which is connected to GPIO GND (pin 6); TRIG, which serves as the output pin; and ECHO, which serves as the input pin. The hand gesture recognition system controls the motor movements, enabling navigation in four directions: Forward, Backward, Left, and Right, as well as a Stop command. The technology has reached a recognition accuracy of 98%. The system's whole cost amounts to around $200, and it has shown efficient functionality in a pristine setting.

In Figures 8.1b and 8.1c depict the details of the user login screen of the implemented mobile application.Once the Scrren opens up we can initially add a route , update a route as path. A Route can be fixed by providing prior Source and destination details on the Mobile.This route is fixed for particular vehicle and the



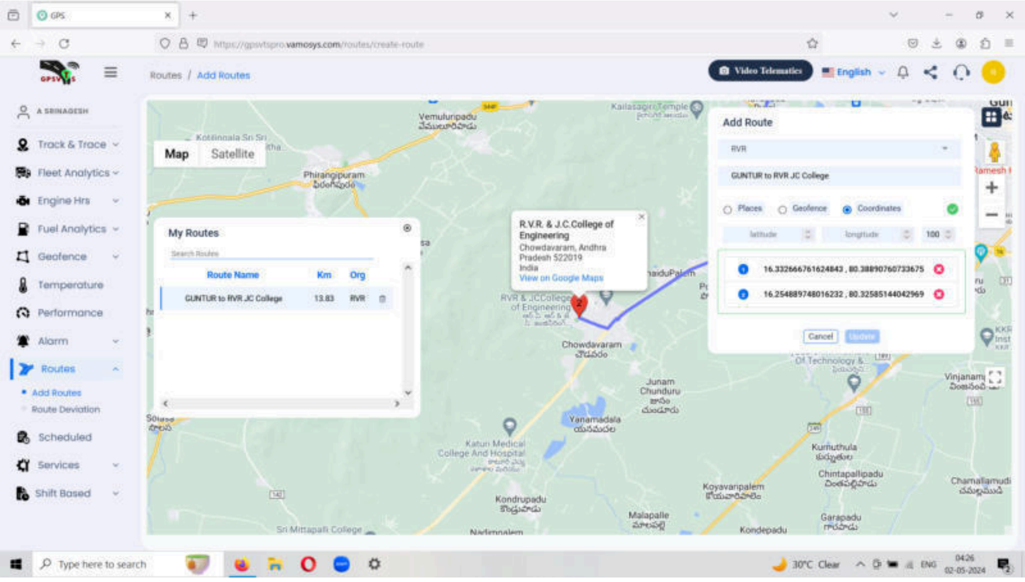
Name: My app
Package name: com.example.myapp
Save location: F:\AndroidStudioProjects\Myapp
Language: Java
Minimum SDK: API 23: Android 6.0 (Marshmallow)
Your app will run on approximately **84.9%** of devices.

(a) Android Mobile Application with navigation details

(b) Login Screen and Authentication

(c) Opening screen of Mobile Application

Fig. 8.1: Screenshots



My Routes

Route Name	Km	Org
GUNTUR to RVR JC College	13.83	RVR

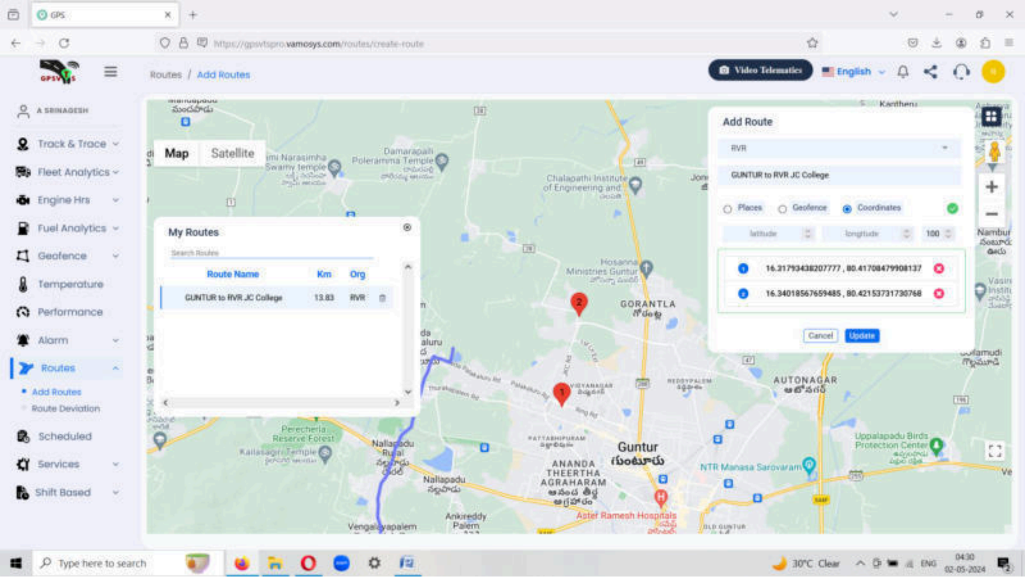
R.V.R. & J.C.College of Engineering
Chowdavaram, Andhra Pradesh 522019
India
View on Google Maps

Add Route
RVR
GUNTUR to RVR JC College
○ Places ○ Geofence ● Coordinates
latitude longitude 100
1 16.332666761624843 , 80.38890760733675
2 16.254889748016232 , 80.32585144042969
Cancel Update

Fig. 8.2: Route Created from Guntur City to RVR & JC College of Engineering

The obtained results are shown in Figueres 8.1b-reff8 and the trip details are can stored in the Firebase using a API key specially generated for a user for a given amount of time to access the Google Maps API.

A legitimate driver is validated with the Authentication Module by confirming his details with the Owner's



My Routes

Route Name	Km	Org
GUNTUR to RVR JC College	13.83	RVR

Add Route

RVR

GUNTUR to RVR JC College

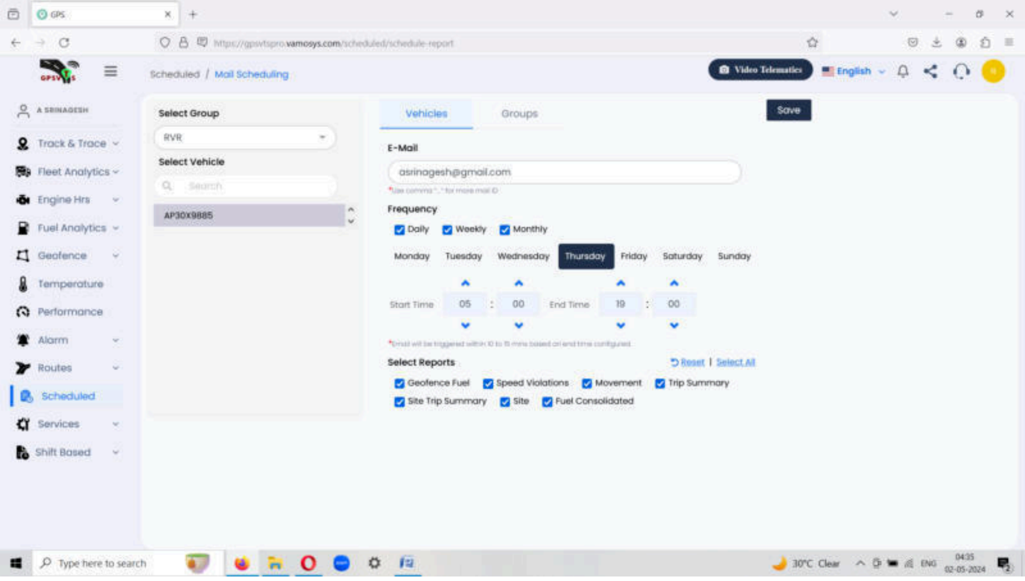
○ Places ○ Geofence ● Coordinates

latitude longitude 100

1 16.21793438207777 , 80.41708479908137

2 16.34018567659485 , 80.42153731730768

Fig. 8.3: Route Update option with Google Maps Coordinates



Select Group: RVR

Select Vehicle: AP30X9885

E-Mail: asrinagesh@gmail.com

Frequency: ☑ Daily ☑ Weekly ☑ Monthly

Monday Tuesday Wednesday **Thursday** Friday Saturday Sunday

Start Time 05 : 00 End Time 19 : 00

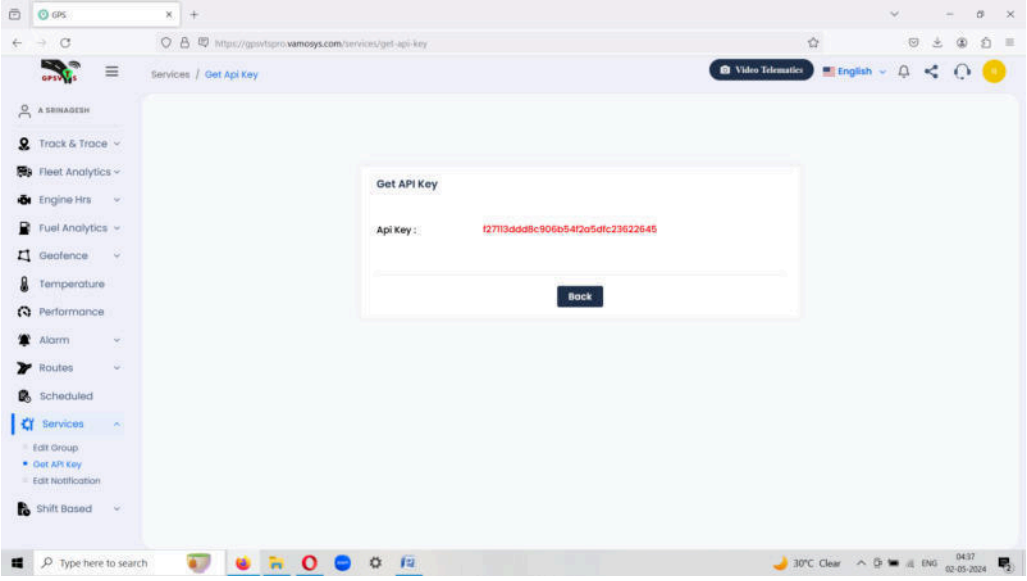
Select Reports

☑ Geofence Fuel ☑ Speed Violations ☑ Movement ☑ Trip Summary

☑ Site Trip Summary ☑ Site ☑ Fuel Consolidated

Fig. 8.4: Details of the trip saved and sent to the Owner

Firebase database.URL Module is implemented to track the live status and send or notify the server to which the system is connected.Sample code of Search Module to get nearby places using Google Maps.It is possible to record the current location of the vehicle in real-time [13]. Using the search module, one can look for a nearby gas bunk. This feature can also be employed in the event of an unexpected sudden incident, such as punctures, or sudden car breakdown. With the search module, one can look for a nearby gas bunk. This feature can also be employed in the event of an unexpected sudden incident, such as a vehicle breaking down, or any form of puncture.



Get API Key

Api Key : f27113ddd8c906b54f2a5dfc23622645

Fig. 8.5: Generate API Key to access the Cloud data

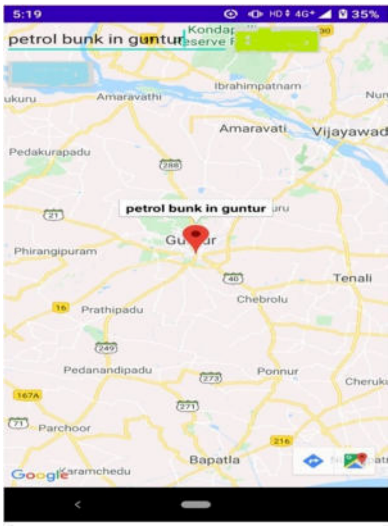


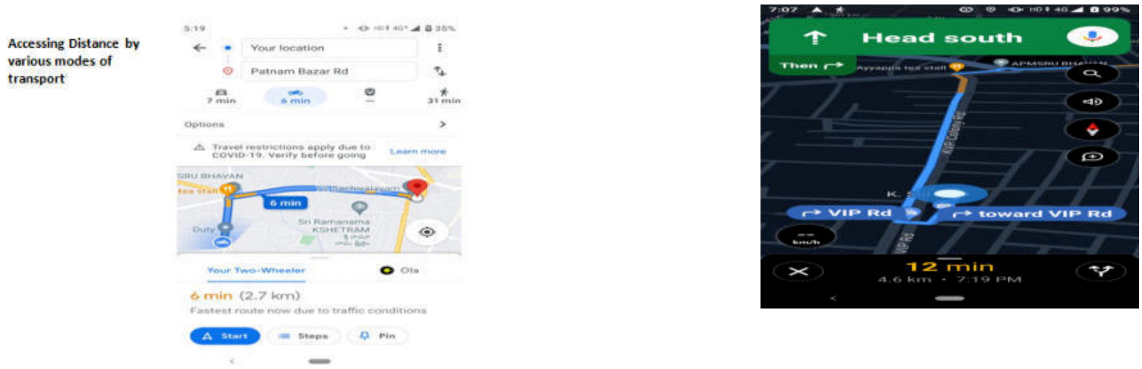
Fig. 8.6: Accessing our location and Searching Petrol bunk option

The details captured in each trip as trip summary by the GPS Fuel Sensor is presented in Table 8.1.

This GPS-based Fuel sensor makes it possible to trace the route information, path deviations, timestamps, and other important details in the Navigation module. The details of the trip summary can be recorded as shown in Table 8.1.

The fuel statistics chart with updated fuel status in the fuel tank (dynamic level-daily) and the distance travelled in kilometres are presented in detail in the above figure. It makes it simple to understand every element of the consumption pattern.

9. Conclusion. This work addresses the pressing issue of fuel consumption and theft in vehicles, which has become increasingly critical in the face of rising fuel costs. The proposed solution employs pre-fitted fuel



(a) A clear description of the Proposed Route using Google Maps

(b) Distance travelled and Route path Information with Navigation details.

Fig. 8.7: Route path

Table 8.1: Sample Fuel Analytics Data

Vehicle Number	AP30X9885
Vehicle Mode	Moving Vehicle
Engine Status	Voltage + Ignition (0.0)
Tank Size (L)	242
Current Fuel (L)	119.39
Minimum Fill (L)	5
Minimum Theft (L)	5
Total Distance (km)	433.43
Start Fuel Level (L)	70.61
End Fuel Level (L)	177.05
Total Fuel Fill (L)	179.11
Total Fuel Theft (L)	0
Total Fuel Consumption (L)	72.67
Kmpl	5.96
Type	Fill
From Time	22-01-2024 08:22:00
To Time	22-01-2024 08:55:25
Previous Fuel (L)	62.89
Current Fuel (L)	242
Fuel (L)	179.11
Nearest Location	Agraharam, GNT, Guntur, Andhra Pradesh, India, 522004

sensors and a fleet management system to monitor fuel usage and routes, offering a range of features such as sensor activation during fueling, data storage in a mobile application for tracking and theft prevention, and immediate alerts to vehicle owners based on consumption patterns. The researchers also suggest a low-cost Vehicle Tracking System (VTS) with Fuel Detection that utilizes ultrasound-based fuel sensing to compare consumption with the traveled route, enabling text alerts for the owner. The results obtained demonstrate the effectiveness of the system, with an average distance traveled for March 2024 of 1958.06 Km, total fuel consumed of 368.02 liters, and Liters per hour value of 90.57. These values are presented in Table 10 of

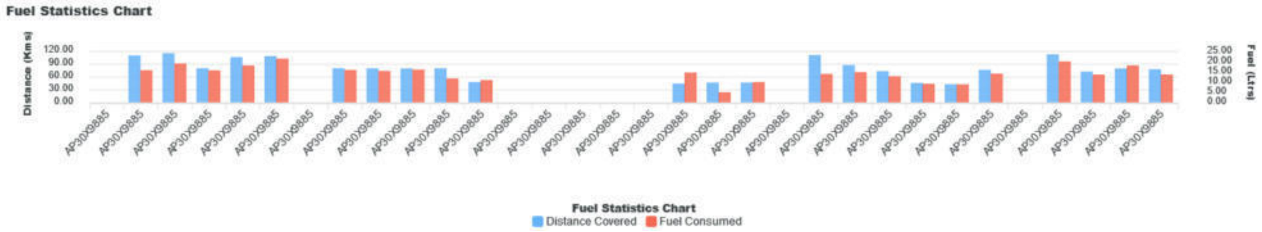


Fig. 8.8: Fuel Statistics chart

Table 8.2: Comparison of Fuel Consumption every month with Kmpl and Filled dates

S.No	Month	Distance (km)	Fuel (L)	Filled Fuel Con- sumed (L)	Kmpl	No. Fills	of Fuel (L) Date	Filled Analysis with				
1	April-2023	1919.86	172.86	346.88	5.53	1	172.86					
2	May-2023	1467.27	346.35	270.84	5.42	3	185.03 2023]	[01-05-	102.99 2023]	[12-05-	58.33 2023]	[25-05-
3	June-2023	1750.86	273.25	327.94	5.34	2	109.17 2023]	[07-06-	164.08 2023]	[20-06-		
4	July-2023	2028.12	363.92	373.73	5.43	2	178.89 2023]	[03-07-	185.03 2023]	[15-07-		
5	August-2023	1767.76	357.80	340.06	5.2	3	173.89 2023]	[03-08-	128.9 2023]	[21-08-	55.01 2023]	[30-08-
6	September-2023	2375.47	443.82	455.58	5.21	3	160.51 2023]	[01-09-	127.2 2023]	[09-09-	156.11 2023]	[22-09-
7	October-2023	1999.14	341.55	359.15	5.57	2	158.49 2023]	[03-10-	183.06 2023]	[13-10-		
8	November-2023	2524.81	473.79	450.38	5.61	3	175.11 2023]	[01-11-	148.94 2023]	[10-11-	154.14 2023]	[22-11-
9	December-2023	2015.59	363.07	377.34	5.34	3	182.61 2023]	[02-12-	132.77 2023]	[13-12-	47.69 2023]	[28-12-
10	January-2024	1714.17	351.68	311.94	5.5	2	172.57 2024]	[02-01-	179.11 2024]	[22-01-		
11	February-2024											
12	March-2024											

Table 8.3: Sample GPS Analytics Report - RVR (02-05-2024 05:01)

S.No	Vehicle Name	Engine Mode	Moving	Parked	Idle	No Data	Stoppage	Distance (Km)	Avg Speed (Km/h)
1	AP30X9885	Voltage + Ignition (0.0)	68h:34m	695h:23m	21h:9m	6m:0s	716h:32m	1958.06	27

the paper, highlighting the system's ability to address the critical challenges of fuel consumption and theft in vehicles. The GPS-enabled Fuel Sensor-based Vehicle Tracking System for Fleet Management using the Internet of Things is an efficient and comprehensive solution to monitor fuel usage and combat fuel theft, particularly beneficial for fleet management systems. Its implementation can significantly enhance fuel efficiency, reduce costs, and improve overall fleet management.

The system provides real-time monitoring of fuel consumption, route deviations, driving habits, breakdown details, vehicle maintenance conditions, and other vital performance details of a Heavy vehicle, which enables vehicle owners to receive real-time reports on their mobile devices. In addition, the system's real-time analysis of Vehicle Performance and Fuel Analytics can provide valuable insights for vehicle owners. Although the

Table 8.4: Total Distance Covered during March 2024

S.No	Date	Distance	Date	Distance	Date	Distance
1	01-Mar-24	110.96	11-Mar-24	107.56	21-Mar-24	72.48
2	02-Mar-24	111.42	12-Mar-24	78.49	22-Mar-24	106.86
3	03-Mar-24	0	13-Mar-24	23	23-Mar-24	109.32
4	04-Mar-24	111.2	14-Mar-24	74.61	24-Mar-24	0
5	05-Mar-24	109.13	15-Mar-24	71.24	25-Mar-24	0
6	06-Mar-24	109.61	16-Mar-24	72.37	26-Mar-24	72.362
7	07-Mar-24	110.35	17-Mar-24	0	27-Mar-24	74.45
8	08-Mar-24	0	18-Mar-24	74.9	28-Mar-24	105.8
9	09-Mar-24	0	19-Mar-24	75	29-Mar-24	0
10	10-Mar-24	0	20-Mar-24	71.29	30-Mar-24	105.4
			31-Mar-24	0	**Total Distance Covered**	**1958.06 Kms**

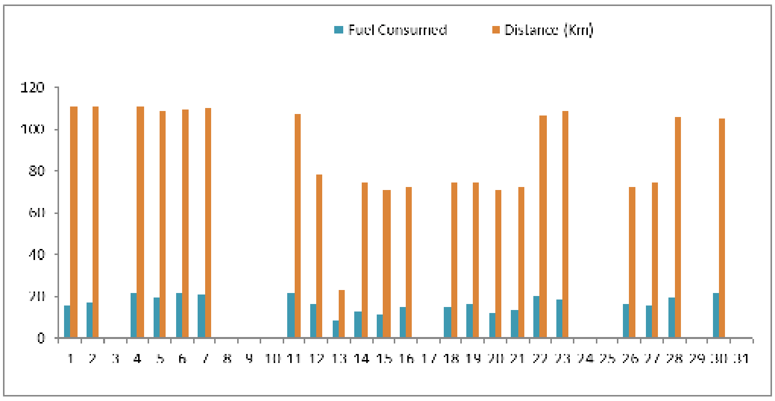


Fig. 8.9: Fuel Consumed and Distance Travelled in March 2024 Statistics chart

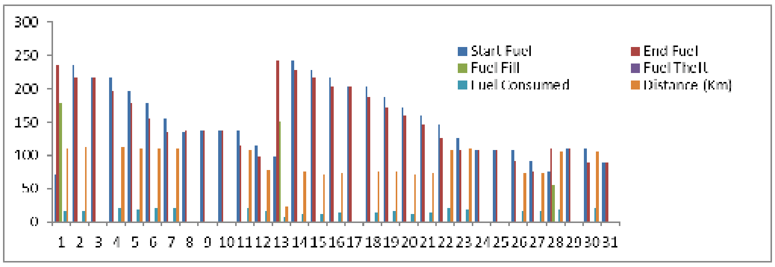


Fig. 8.10: Fuel Details with Distance Travelled in March 2024 Statistics chart

system has some limitations, such as interference factors, it can still be developed and improved in the future to provide even more accurate and comprehensive data analysis. Overall, the GPS-enabled Fuel Sensor-based Vehicle Tracking System for Fleet Management using the Internet of Things is a reliable and effective solution that can benefit vehicle owners and fleet managers in various industries.

Acknowledgments. The resources needed to complete this study were provided by Glocal Technologies Private Limited, Vijayawada, for which the authors are grateful.

Table 8.5: Day-Wise Distance Travelled and Fuel Consumed for the Vehicle for March 2024

S.No	**Fuel Consumed (Litres)**	**Distance (Km)**	**KMPL**	**(L/H)**
1	15.42	110.96	7.2	3.1
2	16.81	111.42	6.6	3.55
3	0	0	0	0
4	21.29	111.2	5.2	4.33
5	18.99	109.13	5.7	4.11
6	21.41	109.61	5.1	4.24
7	20.57	110.35	5.4	4.17
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	21.7	107.56	5	4.37
12	16.09	78.49	4.9	4.2
13	8.18	23	2.8	3.1
14	12.73	74.61	5.9	3.81
15	11.59	71.24	6.1	3.49
16	15.01	72.37	4.8	4.35
17	0	0	0	0
18	14.67	74.9	5.1	4.31
19	16.09	75	4.7	4.53
20	11.85	71.29	6	3.55
21	13.69	72.48	5.3	4.2
22	20.31	106.86	5.3	4.31
23	18.69	109.32	5.8	3.96
24	0	0	0	0
25	0	0	0	0
26	16.32	72.62	4.4	5.4
27	15.76	74.45	4.7	4.66
28	19.21	105.8	5.5	3.96
29	0	0	0	0
30	21.64	105.4	4.9	4.87
31	0	0	0	0
Total	368.02	1958.06	116.4	90.57
Average	11.87	63.16	3.75	3.1

REFERENCES

[1] Chiwhane, S.A., et al. “IOT Based Fuel Monitoring for Future Vehicles”. *International Journal of Advanced Research in Computer and Communication Engineering*, 6, 295-297. (2017)

[2] Padmaja, B.V., et al. “IoT-Based Implementation of Vehicle Monitoring and Tracking System Using Node MCU”. *International Journal of Innovative Technology and Exploring Engineering*, 8, 446-450. (2019)

[3] Dukare, S.S., Patil, D.A. and Rane, K.P. “Vehicle Tracking, Monitoring, and Alerting System: A Review”. *International Journal of Computer Applications*, 119, 39-44. `https://doi.org/10.5120/21107-3835`. (2015)

[4] Gullipalli, S., Karri, Y. and Kota, S. “GPS Live Tracking of Buses and Fuel Monitoring System Using Arduino”. *International Journal for Research in Applied Science & Engineering Technology*, 6, 2278-2285. `https://doi.org/10.22214/ijraset.2018.3362`. (2018)

[5] H. L. Ribeiro and A. Gonzaga, ”Hand Image Segmentation in Video Sequence by GMM: a comparative analysis,” *2006 19th Brazilian Symposium on Computer Graphics and Image Processing*, Amazonas, Brazil, (2006), pp. 357-364, doi: 10.1109/SIBGRAPI.2006.23.

[6] Vanmore, S.V., et al. “Smart Vehicle Tracking Using GPS”. *International Research Journal of Engineering and Technology*, 4, (2017).

[7] Alshamisi, H., and Këpuska, V. “Real-Time GPS Vehicle Tracking System”. *International Journal of Advanced Research in Electronics and Communication Engineering*, 6,(2017).

[8] Rohitaksha, K., Madhu, C.G., Nalini, B.G., and Nirupama, C.V. “Android Application for Vehicle Theft Prevention and

Tracking System". *International Journal of Computer Science and Information Technologies*, 5, (2014).

[9] Khin, M.M. and Oo, N.N. "Real-Time Vehicle Tracking System Using Arduino, GPS, GSM, and Web-Based Technologies". *International Journal of Science and Engineering Applications*, 7, 433-436. (2018), https://doi.org/10.7753/IJSEA0711.1006

[10] Saini, J., Agarwal, M., Gupta, A. and Manjula, R. "Android App Based Vehicle Tracking Using GPS and GSM". *International Journal of Scientific & Technology Research*, 6, 53-58, (2017)

[11] Rasheda Khatun, Sabbir Ahmed Antor, Ahsan Ullah, Afzal Hossain. "Vehicle Fuel Activities Monitoring System Using IoT", *Advances in Internet of Things*, 2019

[12] Kucera, Pavel, Ondrej Hyncica, Petr Honzik, Karel Pavlata, and Karel Horak. "On vehicle data acquisition system", *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011.

[13] N. C. Maduka, M. H. Ibrahim. "Microcontroller-Based Vehicle Tracking System Using GPS and GSM Module: A Mini Review", *International Journal of Science for Global Sustainability*, 2023.

[14] Yoonjong Cho, Jeongwoo Park, Wang-Geun Lee, Jaehyun Park et al. "Prevention of Carbon Corrosion by TiC Formation on Ti Current Collector in Seawater Batteries" , *Advanced Functional Materials*, 2023.

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Mar 11, 2024
Accepted: Jun 10, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 48–63, DOI 10.12694/scpe.v26i1.3501

A PERSPECTIVE STUDY ON SCALABLE COMPUTATION MODEL FOR SKIN CANCER DETECTION: ADVANCEMENTS AND CHALLENGES

N. KAVITHA,[*] N. SIVARAM PRASAD,[†] SUJEETH T,[‡] PALEM NARESH [§] R.SWATHI ,[¶] A.V.L.N SUJITH ,[‖] AND P. DILEEP KUMAR REDDY[**]

Abstract. In the realm of scalable computing, the quest for early detection of skin cancer takes on a new dimension, demanding robust and efficient algorithms capable of handling vast amounts of data. This article delves into the burgeoning field of intelligent computing, where scalable solutions are imperative for processing the multitude of skin lesion images generated daily. Leveraging cutting-edge deep learning and machine learning techniques, researchers strive to develop automated systems capable of swiftly analyzing lesion features like symmetry, color, size, and shape.Through a comprehensive literature review, this paper explores the strides made in skin lesion detection, focusing on scalable computing approaches that accommodate the growing volume of medical imaging data. By identifying significant contributions in classification and segmentation methods, the article not only sheds light on the latest advancements but also offers guidance for aspiring researchers navigating the complexities of skin lesion analysis. Ultimately, the fusion of scalable computing and intelligent algorithms holds promise in revolutionizing early detection efforts, potentially saving countless lives by swiftly identifying and treating skin cancer at its onset.

Key words: Machine learning, deep learning, skin cancer and scalable computing

1. Introduction. One of the most important health problems that the world faces is cancer [1]. As a consequence of the sickness, the human body may exhibit a wide range of distinct symptoms and locations. One of the most common and significant forms of cancer that affects women is breast cancer. Within the male population, prostate cancer is one of the most well-known and fatal forms of cancer. Mesothelioma is a kind of skin cancer that affects both men and women and often results in death. This particular kind of skin cancer is the most common type in the United States, and nine percent of the population is affected by it. In addition, according to the findings of a recent study, the most common cause of death in the United States that is attributed to cancer is melanoma, which causes skin cancer. According to the findings of a recent study, the number of newly diagnosed instances of cancer and deaths attributed to cancer has been assessed.

In the United States of America, skin cancer is one of the most prevalent forms of cancer. Due to the fact that the skin is the largest organ in the body, skin cancer is the kind of cancer that occur most often in people [2]. Two of the most prevalent kinds of skin cancer are melanoma and non-melanoma skin cancer. Melanoma is quite rare. The skin cancer known as melanoma is a rare and potentially lethal form of the disease. Despite the fact that melanoma skin cancer accounts for just one percent of all cases, the American Cancer Society claims that it has a greater prevalence of fatalities [4]. In the cells known as melanocytes, melanoma is able to progress. When healthy melanocytes grow out of control, they transform into cancerous tumors. All areas of the body are susceptible to being affected by it. It is common for individuals to have it on their hands and face since they are constantly exposed to the sun. The only method to cure melanoma cancer is to detect it at an early stage, before it extends to other parts of the body and causes the individual to suffer a horrible death [5]. Melanomas may take many distinct forms, including nodular melanoma, spreading melanomas, and lentigo

[*] Department of CSE, Narsimha Reddy Engineering College, Secunderabad,Telangana State, India (kavitha.chundi@gmail.com)
[†] Department of IT, Bapatla Engineering College, Bapatla, Andhra Pradesh, India.(sivaram.n@becbapatla.ac.in)
[‡] Department of Computer science and Engineering, Siddhartha Educational Academy Group of Institutions, Tirupati, Andhra Pradesh, India (sujeeth.2304@gmail.com)
[§] Department of CSE, Narsimha Reddy Engineering College, Secunderabad, Telangana State, India (nareshpalem09@gmail.com)
[¶] Department of CSE, Sri Venkateswara College of Engineering, Tirupati, Andhra Pradesh, India (swathi.mani08@gmail.com)
[‖] Department of CSE, Narsimha Reddy Engineering College, Secunderabad, Telangana State, India (sujeeth.avln@gmail.com)
[**] Department of CSE, Narsimha Reddy Engineering College, Secunderabad, Telangana State, India (dileepreddy503@gmail.com)

malignant [3] skin tumors. Melanomas are common in the United States. Squamous cell carcinoma (SCC) and basal cell carcinoma (BCC) are the two kinds of non-melanoma skin cancer that comprise the majority of cases (SGC). BCC and SCC are the most frequent types of skin cancer. When it comes to the epidermis, these three forms of cancer are present in the intermediate and upper layers. A low likelihood of the disease spreading to other parts of the body is associated with this kind of cancer. Melanoma malignancies are much more difficult to treat than non-melanoma tumors, which are far simpler to cure.

An extraordinary increase in the use of artificial intelligence (AI) has been seen over the course of the last decade. These variables have led to a considerable progression in computer technology as well as the construction of new algorithms, as well as a spike in the amount of digital data that has been created. At the moment, artificial intelligence (AI) is at the center of a broad variety of tasks that are performed on a daily basis, and it is becoming an increasingly crucial component of the conveniences that are commonly available today. Because these technologies are starting to have an effect on the economy and industries throughout the world, artificial intelligence has become an essential component of important activities in the fields of engineering, finance, and other fields. When it comes to the classification of skin cancer using computer vision, which is a subset of artificial intelligence, deep learning has allowed artificial intelligence to approach near to the level of a dermatologist based on research studies that have been conducted over the last two years. When it comes to dermatology, on the other hand, the usage of these models has been the subject of experimentation for generations. The purpose of this research is to determine the role that dermatologists play in the creation of these models, with the end goal of describing the growth of artificial intelligence in the diagnosis and assessment of skin cancer.

The most essential aspect in determining a patient's prognosis is the early identification of skin cancer, which is a commonly known fact. When it comes to the identification of skin cancer, specialists often use the biopsy method. For the purpose of diagnosis, samples of skin lesions that are thought to be cancerous are taken and submitted to a pathologist. This is a cumbersome, uncomfortable, and time-consuming operation. In the future, the use of computers may make the process of diagnosing skin cancer far easier and more expedient. A variety of non-invasive therapies are available for the purpose of diagnosing skin cancer. These treatments may be used regardless of whether or not the symptoms are suggestive of melanoma. Steps in the process of detecting skin cancer include collecting a picture, processing it, dividing it into smaller pieces, identifying the characteristic that is pertinent, and lastly determining if the image is benign or malignant.

The field of machine learning has been significantly influenced by deep learning in recent times. For the purpose of learning, the area of artificial neural network algorithms is regarded to be at the forefront of the discipline. Their structure is modeled after the way the human brain processes information. The fields of bioinformatics, pattern identification, and voice recognition are all examples of sectors that have discovered applications for deep learning. The use of deep learning systems has been shown to be more successful than the use of traditional machine learning methodologies in some fields. Over the course of the last several years, a variety of deep learning algorithms have been researched for their potential use in the field of computerized skin cancer diagnosis. The objective of this research is to create methods that make use of deep learning in order to identify skin cancer at an earlier developmental stage.

Artificial neural networks (ANN), convolutional neural networks (CNN), self-organizing neural networks (KNN), and generalized adversarial networks (GANs) are some of the technologies that may be used for the detection of skin cancer. The purpose of this study is to conduct a comprehensive and methodical literature review of the many approaches that may be used to diagnose skin cancer. There has been a significant amount of research conducted on this topic. In light of this, it is of the utmost importance to collect and evaluate the research, classify them, and synthesise the findings of the studies that are already accessible. For the purpose of conducting a comprehensive systematic review of skin cancer detection systems that are based on deep neural network-based classification, we used search strings to retrieve relevant content of interest. Our investigation was concentrated on conferences and publications of a high standard. Following the application of our multi-stage selection criteria and assessment technique, we devised a search strategy that resulted in the discovery of sixty-five articles that were of interest. The works in question were subjected to a comprehensive analysis and evaluation from a wide range of viewpoints. Although there have been some encouraging developments in the detection of skin cancer, there is still space for improvement in the diagnostic procedures that are now in use.

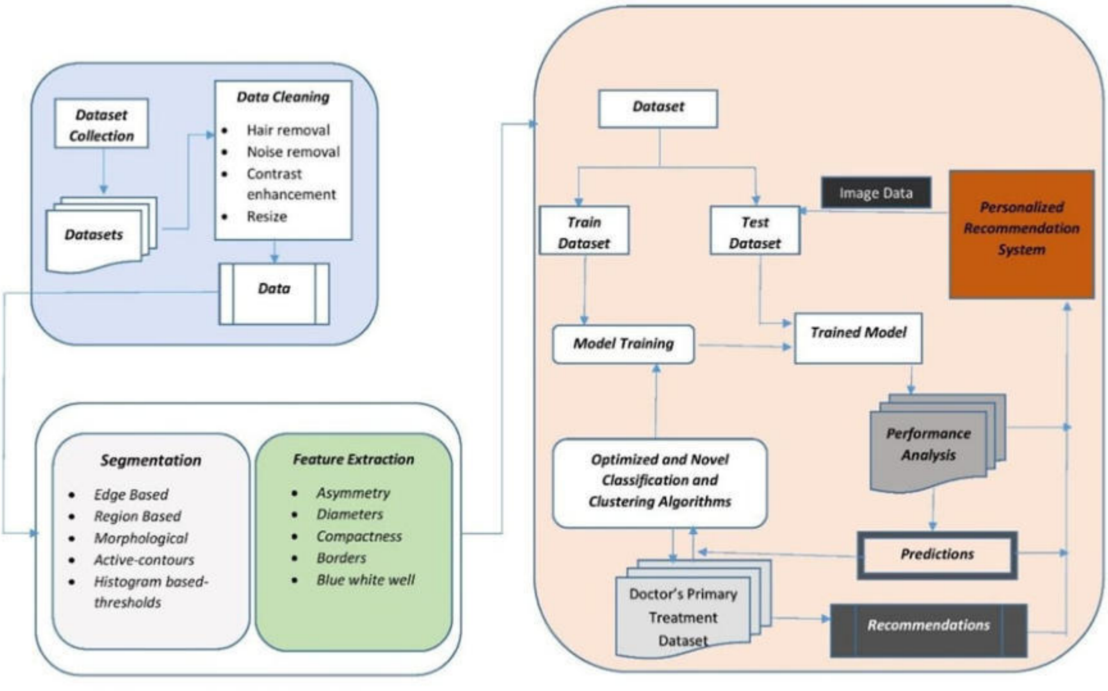


Fig. 1.1: Generic architecture for Early Diagnosis of Skin Cancer Using Machine Learning

1.1. Role of diabetes in Skin Cancer. Individuals who have diabetes mellitus have an increased risk of developing a certain kind of cancer. On a number of times, researchers in Taiwan have focused their attention on individuals who have diabetes mellitus and the risk of developing skin cancer [6]. Using information obtained from the Taiwan Longitudinal Health Insurance Research Database, this retrospective cohort research evaluated the likelihood of acquiring melanoma and non-melanoma skin cancers (NMSCs) between persons who had diabetes and those who did not have diabetes. Patients with diabetes mellitus, who also have an increased tendency for cell proliferation, have raised insulin and IGF (insulin-like growth factor) levels, which results in the production of mitogen and anti-neoplastic effects, as well as malignant cell transformation.

Melanoma is the most common type of skin cancer, followed by squamous cell carcinoma, basal cell carcinoma, malignant tumor of sebaceous glands and sweat glands, and non-melanoma skin cancer (NMSC). Melanoma is the most common form of skin cancer. The incidence of these two malignancies was among the highest ever recorded in Taiwan [7]. In spite of this, the risk of skin cancer in diabetics has gotten a lower amount of money for scholarly investigation. In light of the fact that only a limited number of studies have shown a connection between diabetes and malignant melanoma, it is not obvious whether or not this link is indeed present in other nations.

1.2. Personalized Diagnosis and Early Treatment Recommendation System for Melanoma Patients. There is a possibility that machine learning algorithms may bring about a significant change in the existing way of detecting skin cancer. In order to enhance cancer diagnosis rates, they may concentrate their limited resources on those individuals who are most likely to be affected by the illness. If they were used, patient visits would be simplified, and there would be an increase in the number of referrals to dermatologists. Another possible use for dermatologists is the utilization of mobile apps for the purpose of providing clinical decision help throughout the course of service.It is possible that visual explanations of the qualities that a model uses for classification might also be valuable in diagnosis; with the use of a decision support app, a physician could get a comprehension of both the prediction made by the model and the classification method that it uses. Dermatologists may utilize this information to either narrow down the probable reasons of a patient's symptoms or incorporate it in a full-body skin exam for a more precise diagnosis [8]. This is only one of the many possible applications for this information.

Almost two-thirds of all mobile apps that are relevant to dermatology provide users the ability to monitor their own skin lesions by making use of the camera that is built into their device. with the purpose of providing patients with the ability to engage in teleconsultation with their physicians on any concerns they may have. Users are able to detect lesions, follow diagnostic algorithms, register customized prescription regimens, and record symptoms via the variety of possibilities offered by individualized monitoring programs. These applications also allow users to document symptoms. Users are able to take digital photographs of moles and other lesions via the use of an application called LoveMySkin Mole Map, which was developed by the University of Michigan Medical Center. Additional programs that allow users to capture photographs of moles include FotoSkin, Embarrassing Bodies – My MoleChecker, and UMSkin-Check. You may download these applications from the internet. A three-dimensional model is used by Apre Skin in order to classify and record moles, hence elevating the degree of documentation [9].

Two recent instances of the rapid spread of telemedicine into mobile technology platforms for the delivery of health care are direct-to-patient and patient-directed teledermatology as well as teledermoscopy. This is most likely owing to the introduction of smartphone dermatoscopes that may be used in combination with high-resolution built-in digital cameras and high-speed Internet connectivity. Teledermatology may be able to provide a solution to the problem of patients who are unable to attend dermatologists owing to factors such as transportation, financial restrictions, or time limits. In the present day, around eight to ten percent of direct-to-consumer teledermatology practices are exclusively mobile-based teledermatology services. The teledermatology app known as DermCheck costs a monthly fee of twenty dollars for unrestricted access to the dermatological concierge service. In contrast, other teledermatology services charge anywhere from forty to one hundred dollars for a single appointment. The screening of patients for skin cancer is one of the most prominent applications for these services; however, they may also be used to treat a wide variety of other dermatological conditions as well [10].

Patients have the ability to initiate a consultation with a dermatologist by using a smartphone application that adheres to the principle of direct patient care. Generally speaking, they use a method known as "store-and-forward," in which patients submit their medical histories and digital photographs for the purpose of being evaluated by dermatologists. These dermatologists then offer a consultation and appropriate treatment within a time limit that has been established beforehand. It is possible that you should make use of teledermatology apps if you are worried about suspicious lesions that are present on your skin. A good example of this would be the Skin Cancer App Dermatologist, which offers a consultation within twenty-four hours and costs a fee of nine dollars. An opinion from a dermatologist is inferred from the name of the app, despite the fact that the app itself just specifies that it is an opinion from a "genuine doctor." When compared to face-to-face consultations, the diagnostic accuracy of teledermatology and teledermoscopy varies, however there are certain situations in which they are comparable to one another. According to a recent study, a significant number of direct-to-consumer teledermatology platforms did not identify or certify the consulting physician or hired physicians who were not licensed to practice medicine in the state of the patient or even in the United States. Additionally, it was rare to comply with guidelines in the area of teledermatology, such as providing patients with a choice of experts and presenting a report back to their primary care physician. In the same study, three out of fourteen consultants failed to appropriately identify a nodular melanomas as being concerning. Quite commonly, there is uncertainty about the quality of the service that is being provided. Concerns have also been raised about the lack of privacy safeguards and regulations that are included in mobile teledermatology apps.

When sensitive photos are involved and the treating physician or provider is located outside of the proximity zone, this particular situation is very concerning. The fact that the majority of smartphones destined for the future generation have the capability to monitor personal information, such as the location of the user, makes this situation much more precarious. There are just a few of programs, such as SkyMD and DermEngine, that provide you the opportunity to protect your privacy. It has been determined that a few of these apps do not meet the criteria established by the American Academy of Dermatology for teledermatology of superior quality.

Table 1.1: Types of Intelligent Mobile applications for skin cancer detection

Type of Mobile Application	Pros	Cons
Applications that educate the users about early diagnosis of skin cancer	Inexpensive and most effective way to educate patients about symptoms of skin cancer that helps the individual for early diagnosis	Many of such applications were not backed up with a systematic process to verify the accuracy of the information preloaded in the application
Mole Mapping: This technique enables the app users (patient) to share their images pertaining to area of concern	This technology enables the patient to collect their own images for self-examination	Quality of the image is variable on not suitable for diagnosis
Teledermatology: This mechanism enables a platform for patient-directed virtual treatment	This kind of applications provide access to dermatologists without temporal and geographic barriers	Many of such applications are not adhered to the standards of telemedicine

A relatively recent occurrence in the field of teledermatology is known as patient-directed teledermoscopy. Patients have the ability to take dermatoscopic images of their skin using the MoleScope, a smartphone-mounted dermatoscopy device that costs $99, and then transmit these photographs to a dermatologist for diagnosis. This procedure is not commonly utilized, requires specialized tools, and is most likely best suited for high-risk patients whose dermatologists are knowledgeable with this modality. Despite the fact that research has proved that this method is theoretically practical and acceptable to patients, it is not generally employed.

New ethical concerns need to be addressed if this new technology is to be fully realized while minimizing the amount of harm that is caused to patients. This is because smartphone apps are becoming more widespread. Before the usage of mobile apps can be regarded ethically acceptable, there are a number of issues that need to be addressed, including concerns about the privacy of patients, informed consent, transparency of data ownership, and protection of data privacy. The rapid advancement of this technology has resulted in the construction of a system that is capable of certifying a level of care being beyond its capabilities. Although guidelines for teledermatology have been created, the degree to which these standards are adhered to is still largely up to the discretion of the practitioner.

This paper aims to undertake a comprehensive literature review focusing on technology-enabled options for early diagnosis of skin cancer. Recent research have shown a predominant emphasis on developing advanced machine learning algorithms to detect skin cancer in its early stages. It was noted that only a small number of studies focused on creating a customized recommendation system for early detection and treatment of skin cancer. This study involves a thorough examination of several research articles from different publications to determine the extent and development of research. The article is structured as follows: part 2 explains the research methods used in the study, while section 3 provides an in-depth overview of studies that have significantly contributed to predicting skin cancer in its early stages. Section 4 provides information on the research gaps and potential areas for additional study in skin cancer diagnosis. Section 5 provides the last comments of the study.

2. Research Methodology. The primary goal of this SLR is to provide the groundwork for future studies by outlining the areas of study that need more investigation into the topic of intelligent algorithms for the early detection of skin cancer, as well as any gaps in the current body of knowledge. It has been noted that there is a lack of well-interpreted SLRs addressing the important algorithms involved in creating deep learning algorithms for effective picture analysis, even though there have been numerous published SLRs tackling different obstacles to the development of efficient protocols in technology-enabled cancer treatment. The first step is to establish a review protocol, as shown in figure 2.1. This will allow us to formulate initial Research Questions (RQs) based on a systematic search of over-indexed journal databases using keywords related to the wireless sensor network domain. Our goal is to find recent studies that have addressed this topic and have focused on developing efficient algorithms for computation in sensor networks. Step two of the review methodology involves identifying relevant preliminary studies; step three involves documenting and interpreting the results of the thorough study; and finally, step four involves determining the scope of future research by using the established inclusion and

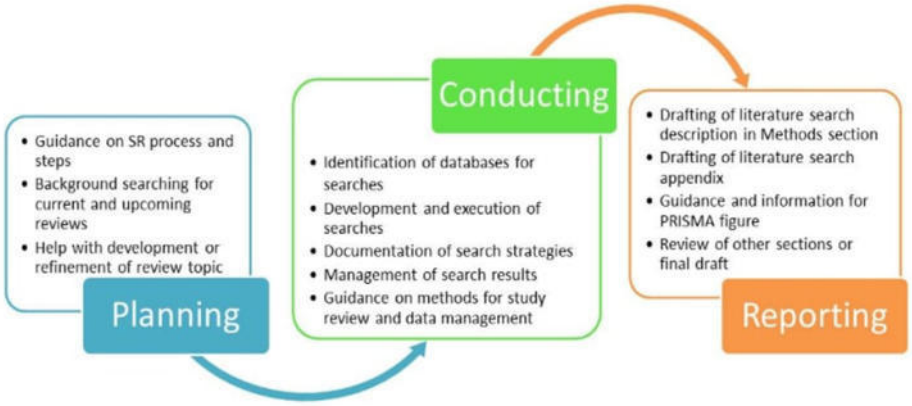

Planning
• Guidance on SR process and steps
• Background searching for current and upcoming reviews
• Help with development or refinement of review topic
Conducting
• Identification of databases for searches
• Development and execution of searches
• Documentation of search strategies
• Management of search results
• Guidance on methods for study review and data management
Reporting
• Drafting of literature search description in Methods section
• Drafting of literature search appendix
• Guidance and information for PRISMA figure
• Review of other sections or final draft

Fig. 2.1: Systematic review process [11]

Table 2.1: Combination of search strings to identify relevant articles from scientific databases

Skin Cancer OR skin Lesion OR Melanoma OR Non Melanoma Skin Cancer
AND
Epidermis OR Hypodermis OR Dermis OR Cancerous tumor
AND
Basal Cell Carcinoma OR Squamous Cell Carcinoma,OR Merkel Cell Cancer
AND
Dermatologist OR Surgical Oncologist OR Hierarchical Based OR radiation oncologist
AND
Computational Intelligence algorithms OR heuristic algorithms OR machine learning techniques OR Meta heuristic algorithms
AND
Personalized Medicine OR Telemedicine OR Mobile Applications OR Biopsy OR MRI OR CT-SCAN
AND
Systematic Study OR SLR OR Mapping Study OR Review

exclusion criteria.

2.1. Search strategy. The search strategy is developed by applying a predetermined set of keywords to indexed databases such as IEEE, ACM, SPINGER, SCIENCE DIRECT, etc. Table 2.1 shows the combination of search strings used for preliminary article search, which are related to computer networks and wireless sensor networks:

In the initial cases based on above search strings 148 relevant research papers addressing the domain of wireless sensor networks were identified within a range of a decade (2011-2021) directly from scientific databases the dissemination of the articles over various databases is depicted in Table 2.2.

2.2. Defining Research Questions. The need and impetus for doing a systematic review are part of the research question formulation process shown in Table 2.3 which is seen as an essential part of analyzing an SLR. Based on the insights provided by [12], the PICO technique is used to craft robust research questions that will generate high-level evidence to back up the review's findings.

2.3. Preliminary selection. First, 148 articles are culled from scientific databases using the search terms given in table 1.1. Next, we check the articles' titles to see whether they address the topic adequately. Second, we used a Table 2.4 to keep track of which articles made the cut and which ones did not.

Subsequently, 58 research articles meeting the inclusion and exclusion criteria were chosen to document the review. It should be mentioned that these articles provide the necessary information to expand the study's

Table 2.2: Dissemination of identified articles over various scientific databases

S. No	Database	No. Of Papers
1	IEEE	55
2	ACM	25
3	TAILOR and FRANCIS	15
4	SPRINGER	27
5	SCIENCE DIRECT	26
Total		148

Table 2.3: Formulation of Research Questions

Acronym	Definition	Motivation	Research question
P	Problem	Gain knowledge related to in-depth analysis of various deep learning based image processing algorithms	RQ1: What are the various deep learning algorithms utilized for the purpose of segmentation of images in the field of medical diagnosis?
I	Intervention	Understand the state of art algorithms involved in developing efficient prediction and accuracy while detecting skin cancer	RQ2: What are the state of art algorithms and dataset evaluated while diagnosing skin cancer at early stages?
C	Comparison	Comparative analysis of existing algorithms to evaluate the variation of prediction accuracy	RQ3: Generate analysis of various existing algorithms and analyze their performance metrics
O	Outcome	Identify open research issues and challenges in technology enabled intelligent diagnosis of skin cancer	RQ4: What is the future scope of research in deep learning enthused skin cancer detection?

Table 2.4: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Articles that Included algorithms centric methods, deep learning architectures and mathematical assertions designed in the context of addressing skin lesion detection	Articles with an ambiguity in the context of the implementation tools and data sets
Articles developed based on the evidential research that is formulated with well defined implementation details and simulation results along with the inclusion of tools and datasets required for the segmentation and prediction metrics relatives to	White papers and lecture nodes regarding published in the context of the architectural perspective of skin cancer
Articles that are primarily implemented in the computer science domain in specific to the areas of machine learning, deep learning and Artificial Intelligence.	Articles that are written in other than the English language.
Articles that are written in the English language.	

scope in relation to skin cancer detection.

3. Review of Various Existing Studies. Melanoma is a very dangerous kind of skin cancer that kills tens of thousands of people every year. If melanoma is to be treated rapidly, the medical community will need to find solutions to a number of novel challenges. Researchers seeking a better treatment for melanoma should sift through literature reviews based on previous studies that were carried out in different eras and places. Conducting a literature assessment of relevant sources is crucial for gaining a better understanding of the research environment. Using objective methods, researchers may begin studying medical image processing and how to cure melanoma using new technology.

In this section, we review the segmentation and classification methods, and we address the problems that came up when reviewing the literature on sample data. Many studies have investigated the possibility of improving medical research by identifying melanoma using deep learning techniques for dermoscopy images.

3.1. Review of Studies on Skin Lesion Segmentation. Various methods for segmenting the lesion region in dermoscopic pictures are described thoroughly in this portion of the paper.

An approach that employs histogram-based clustering estimations and neutrosophic c-means clustering (NCM) for the input dermoscopy photographs was developed by Amira Ashour and colleagues (2018) [13] for successful skin lesion diagnosis. Sort the dermoscopy images' pixels according to neutrosophic criteria first. The HBCE algorithm calculates using h-v and v-h approaches. The public data set from ISIC 2016 is used for the implementation; it contains 379 test photos and 900 training photographs. Effective training and testing based on the availability of ground truth images is necessary since the valuation is done using ISIC 2016 data sets. In comparison to the gold standard NCM method that does not include HBCE, the results of the proposed research are much better.

An automated technique for lesion segmentation using semi-supervised learning is described by Seetharan-iMurugaiyan Jaisakthi et al. (2018) [14]. The method consists of two steps: pre-processing and segmentation. In the pre-processing step, the bi-linear interpolation technique is used to scale the images, and the CLACHE algorithm is used to optimize the images with uneven lighting. After that, the Frangivesselness filter from FMM is used to swap out the pixels that represent hair. In a segmentation procedure, pixels with consistent color and texture characteristics are used to identify lesion zones. Using this method's boundary and region information, which is reinforced by using RGB-based kmeans clustering, we can divide the foreground image into approximately defined ”lesion areas.”

A fresh approach to addressing this problem was suggested by Sahar Sabbaghi et al. [15] in 2018. With its expertise in color assessment, the Quad-Tree system can tell if a melanoma is benign or malignant only by looking at its color. Examining melanomas in this study included using geometric distances and concentratic quartiles. A higher level of contrast between the lesion and background areas is achieved before processing begins. Lesions without color contrast may be treated using morphological treatments like top-hat and bottom-hat surgeries. The hybrid thresholding method divides the segmentation process into two steps, allowing for accurate identification of lesion borders. In order to find core lesions, we adjust the Otsu threshold. Then, we use an adaptive histogram algorithm to make them longer. Based on the characteristics of the ROC curve, it is determined that the SVM classifier performs the best.

Brammya et al. (2018) [16] created a novel meta-heuristic method by using the DHOA-NN approach. The buck's eyesight is five times stronger than a human's, which makes tracking him down much more difficult. Above the horizon, it is difficult to see what is happening. There seems to be hunting going on in the vicinity. The goal function is used to iteratively update positions until the best possible location is discovered. It follows that DHOA-convergent NN outperforms competing methods in terms of performance.

Amira Soudani et al. (2019) [17] presents a segmentation recommender that is built on top of community sourcing and transfer learning. In order to get features from pre-trained architectures such as ResNet50 or VGG16, the convolutional parts are used. The CNN serves as a classifier, while the five nodes that make up the output layer stand in for different segmentation techniques. It is possible to identify local traits from a variety of angles using the two-dimensional structure of dermoscopy photographs. The results back up our prediction that our suggested strategy might lead to a segmentation approach for skin lesion detection.

Examining the CNN architecture, Walker et al. (2019)[18] demonstrates the usage of the inception v2 network for dermoscopy image classification as benign or malignant. A technique known as stochastic decent gradient is used for training the inception v2 parameters inside the deep learning framework. Among the many possible impacts that dermoscopy pictures might bring about are visual characteristics and sonification. The research demonstrates that tele-dermoscopy imaging has a very sensitive malignancy detector and enhanced accuracy in both pigmented and non-pigmented lesions.

Teck Yan Tan and colleagues (2018) [19] created the Particle Swarm Optimization (PSO) method, which is used to identify skin cancer using dermoscopy photographs. The method's stages of operation involve tasks such as segmenting skin lesions and extracting features, optimizing features based on PSO, and performing classification. Swarm leaders divide the initial population in half and then guide each half to choose the best possible solution while avoiding the worst. This approach lessens the possibility that a PSO model converges too quickly by using international and domestic food and enemy signals, attraction, and mutation-based exploitation. Subswarm leaders are bolstered using random walks such as Gaussian, Cauchy, and Levy. There is a plethora of searchability provided by probability distribution and dynamic matrix representation. The proposed method has enhanced melanoma classification accuracy and has performed well on benchmark problems with either a uni- or multi-modal structure. If you want further evidence of how great the proposed method is, you may use the Wilcoxon rank sum test.

According to Mohammed, Al-Masni et al. (2018) [20] , a new approach of segmenting dermoscopy pictures based on a fully resolution convolution network may investigate the full resolution characteristics of every pixel in the input image. The cross entropy loss function is employed by CNNs for pixel-wise categorization. There are no prior or post processing steps required in the suggested method for obtaining full resolution images. On the network levels, back propagation is used to reduce training error. FrCN outperforms the most recent deep learning segmentation methods in comparison to two publicly accessible datasets, the ISBI 2017 Challenge and PH2 datasets.

Image segmentation approaches are compared by Anuj Kumar and his colleagues (2018) [21]. Analysis and identification of relevant characteristics or objects within an image is known as segmentation. An essential component of image analysis is edge-based segmentation, which displays discontinuities of the edges in terms of intensity. An image threshold value is calculated and then compared to the pixel value in order to eliminate broken edges using the canny edge detector. That an edge exists is based on the higher pixel value. There must be a boundary around the region chosen for segmentation based on region. First, a pre-processing phase minimises noise while keeping the picture information that allows them to get a well-segmented image. Therefore, it can be concluded that the edge detector canny delivers the best performance when employing region expanding, which speeds up the segmentation process when compared to region splitting and merging.

A novel approach to handling CNN variable tuning for process fine-tuning was proposed by Guotai Wang and colleagues (2018) [22] and is called Bounding box and Image Tuning-based Segmentation (BIFSeg). Intuitive 2D and 3D medical picture segmentation is achieved via a deep learning-based approach. The proposed weighted loss function in this method lends credence to both supervised and unsupervised image-tuning. The information in the bounding box teaches a convolutional neural network (CNN) to generalize hidden objects by learning common structures like saliency, contrast, and hyper-intensity across different objects. As a result, compared to prior interactive segmentation methods, the suggested framework BIFSeg increases accuracy while decreasing the amount of time and effort required from the user.

Due to the fine-grained nature of the primary diagnosis of melanoma that may be achieved by early screening and further dermoscopic investigation, such as biopsy and histological evaluation, automated categorization of lesion pictures is a tough issue. As inputs, pixels and illness labels are taken into account while training a single CNN for skin lesion classification. According to this, the AI can categorise skin cancer with improved accuracy when compared to dermatologists, and so achieves higher performance in the detection of most frequent malignancies and the worst forms of skin cancer.

Qaisar et al. (2011) proposed unsupervised segmentation of multiple lesions and improved region-based active contours [23]. Additionally, a level has been automatically chosen using the iterative thresholding approach. Another factor that has helped keep the curves stable is the application of smoothing constraints to the Courant-Friedreichs-Lewy function. In order to assess their method, 320 dermoscopy images of the skin were examined. Their segmentation results, genuine detection rates, and false positive rates have all been enhanced.

In their 2017 publication, Euijoon Ahn and colleagues described a computer-assisted diagnostic method for automatically identifying melanoma by lesion segmentation [24]. Poor skin lesion segmentation performance is caused by a number of challenges with standard segmentation algorithms. These include unclear lesion borders, low contrast between the lesion and surrounding skin, and the lesion touching the image bounds. Improved lesion categorization from surrounding skin regions is attainable using methods derived from sparse representation models and novel background detection algorithms. According to the proposed Bayesian framework, lesions are better described. We validate our approach by comparing it to various traditional lesion segmentation techniques and unsupervised saliency detection methods, based on a comparison of two public datasets. That being said, our method is superior than the others. Applying a saliency-optimization approach might further enhance lesion segmentation.

An algorithmic strategy called computational approach was developed by Roberta et al. (2016) [25] to detect skin lesion sorts based on assessment of the attributes collected from photographs. Their plan includes using a support vector machine, an anisotropic diffusion filter, and an active contour model devoid of edges. Researchers have used many techniques to segment and categorize skin lesions. In order to segment skin lesions in a skin image, Eliezer and Jacob (2016) [26] presented a feature learning method that finds the most essential parts of the image. An innovative method for learning from dictionaries without human supervision called Unsupervised Information-Theoretic Dictionary It was explained how learning works and how it has been used to the segmentation of skin lesions. Results from this research demonstrate that the proposed approach is generalizable to other image segmentation problems.

Andrea et al. (2016) [27] introduced a fast and completely automated way to segment skin lesions in dermoscopy pictures. A training phase was omitted from the application of the Delaunay Triangulation to the skin lesion mask extraction process. Several research have been conducted using the public photo databases.

The co-segmentation methodology was introduced in a study by Leonardo et al. (2017) [28]. It is a new method for segmenting MR images that combines the segmented Biological Target Volume with the segmented Gross Target Volume. Jessica and Filipe (2017) offered fuzzy values that were used to construct a new segmentation algorithm called the melanoma segmentation algorithm. They put their method to the test using 571 photographs taken from the standard ISDI dataset; among them, 446 showed benign skin lesions and 125 depicted malignant melanoma. Their method fared better than the existing algorithms when measured using metrics including balanced accuracy, sensitivity, and Jaccard index. They found that fuzzy-value segmentation was the most successful method out of the bunch.

Hamidi et al. (2017) [29] created a new method for automatic image segmentation by combining saliency with the Otsu threshold approach. Their system, which took skin type and other factors into account, generated a color saliency map and a skin feature saliency map. In addition, a more accurate skin picture was produced by combining the two saliency maps. In order to get more accurate skin lesion borders using the histogram distribution of the pictures, their segmentation approach used a new optimization function to change the usual Otsu threshold strategy. By implementing their algorithm, they demonstrated its superiority in terms of effectiveness and accuracy. Their novel algorithm outperforms and is more robust than existing methods, according to the results of their testing.

Mohamed et al. (2018) [30] created a new way to segment data using full-resolution convolutional networks. The suggested method does not need any pre- or post-processing steps for things like artefact removal, low contrast adjustment, or further enhancement of segmented skin lesion boundaries. An evaluation of the proposed method was conducted using the PH2 dataset and the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 Challenge, both of which are publically available.

Using these highly discriminative qualities in a new segmentation algorithm and a skin lesion detection search method was proposed by Neda and Babak (2018) [31]. Also, a novel two-component speed function was used to carry out the segmentation method utilizing the contour propagation technique. Based on data collected from the skin lesion's periphery, they further included a fresh set of features. These photographs were used to map the peripheral areas of the skin lesions to log-polar space using an existing Daugman's transformation. A variety of features were then extracted from these pictures. They discovered that both the new features used and the linear Support Vector Machine (SVM) for melanomat classification were successful in differentiating melanoma from normal nevus, as compared to other techniques that utilize the existing RUSBoost classification algorithm. To determine which characteristics are most useful for classification, each classifier employs a sequential feature selection method.

3.2. Review of Studies on Detection of Melanoma using Medical Dermasocpic Images. Local Directional Patterns, Local Binary Patterns, and Convolutional Neural Networks are among the feature extraction strategies proposed by Manjunath Rao et al. (2020) [32] and processed by an SVM classifier for effective learning of melanoma lesion pictures. High levels of skin exposure to ultraviolet (UV) radiation are a major contributor to the development of melanoma. Consider both the melanoma and non-melanoma photos for assessment. The SVM classifier is used to classify the three extraction methods. Consequently, the LBP system's classification using the SVM classifier coupled with the polynomial kernel function is more accurate. Advanced LBP might be developed in the future to identify melanoma at an early stage.

There are three high level characteristics that are crucial for the diagnosis of malignant melanoma, and these features are simplified by adding the human interpretation of data on the suggested features, as discussed in Vikash Yadav et al. (2018) [33]. Detection at an early stage is critical, since melanoma is a deadly form of skin cancer that is exacerbated by sun exposure and pre-malignant moles. Comparing the high-level asymmetric features to the low-level asymmetric ones, the suggested high-level features perform well for concave borders. Skin cancer detection and classification may be improved by developing them as an additional tool.

Melanoma may be detected at an early stage with the use of a computer-aided detection (CAD) system, which lowers death rates. Preprocessing, lesion segmentation, feature extraction, feature selection, and classification are all parts of the method. The DullRazor is used to remove the hair and pre-process the lesion, making it easier to see. A more precise segmentation of the lesion is achieved by combining the innovative uniform distribution approach with the active contour method and using the additive rule of probability. It is possible to identify the form and appearance of the lesion using the local intensity gradients distribution by using HOG features derived from the colour, texture and HOG features. It's hard to argue with the improved accuracy and efficiency of the newly proposed diagnostic procedures. PH2 is a publicly accessible dataset including 200 photos that dramatically outperforms previous algorithms when compared. As a consequence, it can be stated that the use of SVM in conjunction with the Boltzman Entropy technique offers better results on entropy-based characteristics.

Electronic shaving (E-Shaver) is an improved technique for removing hair from dermoscopic pictures than the Dull Razor, according to the work of Kiani and colleagues (2011) [34]. The detection of hair direction in the skin using random transform and subsequent Prewitt filters is critical for an effective hair removal procedure. Dermoscopic pictures may be improved by using average thresholding and smoothing to eliminate noise and non-hair features. So it is regarded to be the quickest and easiest method for hair removal that works.

Multi-level feature extraction may be performed by executing decomposition and segmentation effectively, according to a new technique by Sina Khakabi et al. (2012) [35]. There is no need for pre-processing for colour uniformity or artefact removal with this method. Using spatial and colour data, the development pattern of the lesion may be obtained. Tree-based depiction of lesion development patterns is generated by matching each pixel sub-cluster to an individual node in a tree structure. An extensive feature set is made possible by the model's capacity to extract information from various levels of the tree structure. As a result, it's considered an effective framework for extracting features and training models for accurate lesion segmentation.

By analysing images, Omkar Shridhar Murumdar et al. (2015) [36] provide a non-invasive approach that is crucial for diagnostic purposes. Image analysis may provide insight into the lesion's ambiguous information. Steps 1 and 2 form the basis of the proposed system. The Otsu thresholding segmentation delivers excellent results since it is unsupervised. Second, the feature extraction tool, which is the second phase, may be used to evaluate and study photographs appropriately without requiring any type of invasion into the human body. Dermoscopic pictures are processed using a technology to extract the ABCD rule, which identifies the characteristics as asymmetry, border structure, colour variation, and lesion diameter. The TDV value is computed using the values derived from features. The greater the TDV number, the more likely it is that melanoma is present.

Using the ABCD rule, Sharmin Majumder and her colleagues (2018) [37] have devised a framework for determining whether a lesion is malignant or benign. The existence of hair in skin photographs is regarded to be the most difficult duty, even if the differences between malignant and premalignant photos are aesthetically comparable to a greater degree. The difference between the highest and lowest Feret diameters of the best fit ellipse to the skin lesion is used to extract additional features. In the suggested technique, a Back-propagation Neural Network is employed as the classifier (BNN). In the suggested technique, the weights evaluated are the same for all photos and demonstrated to be accurate for all types of photographs.

3.3. Review of Studies on early diagnosis of Skin Cancer using Machine Learning Techniques. Neural networks play a crucial role in the detection of skin cancer. Their structure is built upon interconnected nodes. Their architecture is quite similar to the human brain in terms of the connections between neurons. In order to address specific problems, their nodes collaborate. Once trained, neural networks can execute a certain task at a very high level. We trained neural networks to classify images and identify various skin cancers as part of our study. Many different types of skin lesions are part of the ISIC collection.

According to Xie et al. [38], skin lesions may be classified as either benign or malignant. The proposed system consisted of three parts. The first step in lesion detection in images was to use a self-generating neural network. In the second portion of the investigation, details such as the tumor's border, texture, and color were collected. In all, the system was able to obtain 57 features, including 7 additional ones associated with the description of lesion borders. Principal component analysis (PCA) was used to minimize the feature dimensionality, which led to the selection of the optimal set of features. The last stage was to classify lesions using a NN ensemble model. Ensemble NN may benefit from the use of fuzzy NN and backpropagation (BP) NN for better classification results. Various classification algorithms, including KNN and Adaboot, were also compared to the results of the proposed system. With an accuracy rate of 91.11%, the proposed model outperformed the other classifiers by 7.5% in terms of sensitivity.

A method for automated skin cancer detection based on artificial neural networks (ANNs) was proposed by Masood et al. [39]. The article examined the performance of three artificial neural network (ANN) learning techniques: Levenberg-Marquardt (LM), robust backpropagation, and scaled conjugate gradient. A sensitivity of 92.6% for benign lesions and a specificity of 95.1% for malignant lesions were achieved by SCG learning with a doubling of the number of epochs used. We created a mole classification system to help find skin cancer early on [40]. While extracting features, the proposed method adhered to the ABCD rule of lesions. In the ABCD model, a mole is defined by its form, borders, color, and diameter. A mole's asymmetry and borders were evaluated using the Mumford-Shah algorithm and the Harris Stephen method. Under the new approach, any mole that wasn't black, brown, or cinnamon was considered melanoma. Moles that may be malignant melanoma usually have diameters more than six millimeters (mm). With a backpropagation feed-forward ANN, we were able to classify moles with an accuracy of 97.51%. One possible approach to skin cancer diagnostics is an ANN-based backpropagation system [41]. A 2D-wavelet transform was used by this system for the purpose of feature extraction. The proposed ANN model was used to classify the images as either benign or cancerous. An further method for diagnosing skin cancer using ANNs was proposed by Choudhari and Biday [42]. To segment the pictures, an entropy thresholding technique was used. The skin lesion data was analyzed via a gray-level co-occurrence matrix (GLCM). Photos of skin cancer were accurately classified as either malignant or benign by using feedforward neural networks; the resulting accuracy rate was 86.66%.

According to Aswin and coworkers [43], genetic algorithms (GAs) and artificial neural networks (ANNs) may be used to identify skin cancer. The Otsu thresholding approach was used to extract the region of interest (ROI) from medical imaging software called Dull-Rozar. The segmented pictures were then processed using the GLCM method to extract their distinctive properties. For the categorization of lesion pictures into malignant and noncancerous classifications, a hybrid ANN and GA classifier was utilised.

Using digital dermoscopy pictures, Fengying et al. (2016) [44] came up with an innovative method for determining whether melanocytic tumours are benign or malignant. A self-generating neural network extracts skin lesions, and picture attributes that describe tumour colour, texture, and boundary are also retrieved and identified using a neural network ensemble classifier to classify skin lesions. Dermoscopy skin lesion images are too small in the critical medical context for bigger skin lesions. New border feature methods for assessing border abnormalities across the full and partial lesions have been presented by authors to address this challenging presentation. In their novel technique, an ensemble-based classification algorithm has been devised that blends the normal back propagation neural networks with the fuzzy rules known as fuzzy neural networks in order to achieve improved classification precision. In order to test the effectiveness of their method, they ran a series of tests on two different dermoscopy datasets, which included photos of xanthous and caucasian races.

An expert system developed by Suleiman & Akio (2018) [45] is able to identify the presence of skin cancer from simple photographic photographs of diseased skin patches. The ABCDEs rule has been used to identify melanoma photos in their system. The GrabCut algorithm was used to accomplish the segmentation of an input melanoma picture into skin lesions, and image processing techniques were used to extract characteristics such as the shape feature, colour feature, and geometry feature. Furthermore, the support vector machine and the Gaussian radial basis kernel were used to classify all of the collected characteristics as either malignant or non-cancerous. Melanoma and benign photos have also been used in the different tests. At the conclusion of the research, only six characteristics were shown to be useful in identifying melanoma. In [46,47] the authors have been used machine learning techniques to classify the images and extracting the features.

Melanoma classification may now be done utilising the structural co-occurrence matrix of the major frequencies collected from normal dermoscopy pictures, according to Pedro and colleagues (2018). They've improved their classifying abilities. Researchers from Hongming et al. (2018) [48] reported a computer-aided approach for identifying melanocytic tumours from skin scans. Four modules are included in the suggested technique, as well. Finally, a multi-class support vector machine containing extracted epidermis and dermis characteristics was used to classify the skin picture into several categories, such as melanoma, nevus, or normal tissue. When 66 skin cancer photos are used, their experimental findings show that their model delivers a classification accuracy of more than 95% when used.

4. Observations from the Literature Study. This study aims to identify the most common methods used to diagnose and treat malignant melanoma, a deadly form of skin cancer that may metastasis (spread to other parts of the body). The dermoscopic imaging device may magnify lesions, but its complex design makes it hard to visually inspect. Possible resolution to the problem awaits the implementation of an automated system for skin lesion segmentation and a clustering approach. The concept that precancerous moles and sun exposure may develop into melanoma, a kind of skin cancer, is well knowledge. In order to effectively detect melanoma, the literature study found that segmentation and classification algorithms, with or without pre-processing stages, may be applied.Various feature extraction approaches may be used to obtain these properties from the segmented area of the lesion. It is possible to diagnose melanoma.In order to diagnose skin cancer, the majority of the applications rely on dermoscopyphotographs.Unlike their predecessors, most modern dermatologists rely on manual pattern recognition to detect lesions, drawing on their prior knowledge and experience. Using the dermatoscope to extract features from the lesion, also known as ABCD, allows for an accurate diagnosis. It is possible to use K-means clustering to group together the foreground picture's RGB color space, which has high Jacquard indexes and dice coefficients. Convolutional Neural Networks (CNNs) with weight sharing perform well in image-based skin cancer detection, but they are computationally expensive to train and experience noticeable slowdowns when presented with a huge volume of input dermoscopy photographs. Training data sets using the Support Vector Machine (SVM) approach takes a long time, which is one of its downsides. Using training data to discover new features in the problematic lesion area increases the likelihood that the deep learning approach will provide high-quality results. The optimization techniques used in these real-time engineering applications are tested by comparing their results to those of other state-of-the-art optimization algorithms.

Fuzzy logic-based clustering, pattern classification, image segmentation, fuzzy classification, fuzzy logic under time constraints, and classification have all been extensively studied by several researchers in the past. Nevertheless, no approach has shown to be more effective in reliably identifying melanoma images. The melanoma skin lesion photographs show an improvement in all image segmentation, grouping, and classification procedures.

There were many different approaches to learning deep learning, such as neural network and hybrid methods of fully convolutional neural networks, transfer learning, and ensemble approaches. Both automated deep learning algorithms and more human-centered approaches have shown promising outcomes in the detection of melanoma. The number of images that can be used for training and testing is restricted since most datasets are rather small. The proposed methods reliably provide surprising results when tested on large datasets, although over fitting might be an issue when used to smaller datasets. For instance, there are just 200 images in the PH2 dataset. One possible solution to the problem of training with a small dataset is to use an adversarial generative network in conjunction with data augmentation and transfer learning. Some researchers utilize private datasets and images found online. Since these studies and their findings are not available in a replicable format, and since images seen online can be biased, it would be difficult to reproduce them.

4.1. Open Research Challenges. The extensive training required is a major drawback of skin cancer detection methods based on neural networks. For this reason, getting the system trained to accurately assess and understand the features of dermoscopy images is a laborious and resource-intensive process. A further complicating factor is the fact that lesion sizes may vary greatly. An international group of researchers from Italy and Austria took countless pictures of skin lesions, both benign and cancerous, throughout the 1990s. Diagnosis accuracy in locating lesions varied between 95% and 96%. Diagnosing smaller lesions, those measuring just 1 or 2 millimeters in diameter, at an early stage was much more challenging and prone to errors.

Regular dermoscopy databases are dominated by images of fair-skinned people from Western Europe, Oceania, and the Americas. In order for a neural network to correctly detect skin cancer in people with dark skin, it has to be taught to take skin color into consideration. But this can only happen if black people's faces are used to train the neural network. In order to train skin cancer detection algorithms to be more accurate, datasets should include a sufficient number of images of lesions on people with light and dark skin tones.

There is a significant disparity in the databases used for skin cancer diagnosis in real life. Unbalanced datasets include wildly different amounts of images for each kind of skin cancer. Because of the small sample size of the more uncommon skin malignancies seen in dermoscopy images, it is challenging to draw broad conclusions about the disease based on these images alone. Neural network (NN) software requires robust hardware resources with strong GPU capabilities to extract particular lesion morphological features from images. Inadequate processing power hinders deep learning-based skin cancer detection training. Melanoma risk factors that have been discovered by researchers include a pale complexion, light-colored eyes, red hair, and many moles on the body. When both hereditary and environmental variables are included, the chances of developing skin cancer increase dramatically. When combined with existing deep learning approaches, these features have the potential to improve performance.

5. Conclusion. Methods for identifying and categorizing skin cancers have been investigated in this comprehensive study. Using any of these techniques will not put you in danger. Picture segmentation and preprocessing are prerequisites for skin cancer diagnosis, which include feature extraction and classification. Classification of lesion images using ANNs, CNNs, KNNs, and RBFNs is the main aspect of this study. You can't have an algorithm without its drawbacks. Selecting an appropriate classification scheme is crucial for optimal outcomes. Since CNNs are more often linked with computer vision, they significantly outperform other methods when it comes to recognizing image data. Many skin cancer detection research focus on determining whether a certain image of a lesion is cancerous. Unfortunately, patients sometimes wonder whether a certain skin cancer symptom appears elsewhere on the body, and unfortunately, current research does not provide an answer to this issue. Classifying the signal image has been the only focus of the investigation so far. One possible solution to this prevalent question might be to use full-body photography in future studies. Speeding up and automating the process of image capture is automated full-body photography.

A relatively new idea in deep learning is self-organization. It is an example of unsupervised learning that uses the dataset's image samples to look for patterns and correlations. Expert systems may improve their feature retrieval with the use of convolutional neural networks that employ auto-organization strategies. Currently, there is an active research and development effort centered on auto-organization. Improving image-processing systems for the future, when pinpoint diagnosis of disease depends on paying great attention to the smallest features in medical imaging, may need a better examination of these aspects now.

Acknowledgments. The author thanks the anonymous authors whose work largely constitutes this sample file. He also thanks the INFO-TeX mailing list for the valuable indirect assistance he received.

REFERENCES

[1] Ashraf, R.; Afzal, S.; Rehman, A.U.; Gul, S.; Baber, J.; Bakhtyar, M.; Mehmood, I.; Song, O.Y.; Maqsood, M.,*Region-of-Interest Based Transfer Learning Assisted Framework for Skin Cancer Detection.* IEEE Access 2020, 8, 147858–147871. [CrossRef]

[2] Byrd, A.L.; Belkaid, Y.; Segre, J.A. ,*The Human Skin Microbiome.* Nat. Rev. Microbiol. 2018, 16, 143–155. [CrossRef]

[3] Elgamal, M. ,*Automatic Skin Cancer Images Classification.* IJACSA 2013, 4. [CrossRef]

[4] *Key Statistics for Melanoma Skin Cancer. Am. Cancer Soc.* Available online: https://www.cancer.org/content/dam/CRC/PDF/ Public/8823.00.pdf (accessed on 8 February 2021).

[5] Khan, M.Q.; Hussain, A.; Rehman, S.U.; Khan, U.; Maqsood, M.; Mehmood, K.; Khan, M.A. ,*Classification of Melanoma and Nevus in Digital Images for Diagnosis of Skin Cancer.* IEEE Access 2019, 7, 90132–90144. [CrossRef].

[6] Tseng, H. W., Shiue, Y. L., Tsai, K. W., Huang, W. C., Tang, P. L., & Lam, H. C. (2016).,*Risk of skin cancer in patients with diabetes mellitus: A nationwide retrospective cohort study in Taiwan.* Medicine, 95(26), e4070.

[7] Chen CJ, You SL, Lin LH, et al.,*Cancer epidemiology and control in Taiwan: a brief review.* Jpn J Clin Oncol 2002; 32 (suppl):S66–81

[8] Chao E, Meenan CK, Ferris LK. ,*Smartphone-Based Applications for Skin Monitoring and Melanoma Detection.* Dermatol Clin. 2017 Oct;35(4):551-557. doi: 10.1016/j.det.2017.06.014. Epub 2017 Aug 9. PMID: 28886812.

[9] Kassianos AP, Emery JD, Murchie P, et al. ,*Smartphoneapplications for melanoma detection by community,patient and generalist clinician users: a review.* Br J Dermatol 2015;172(6):1507–18.

[10] Wu X, Oliveria SA, Yagerman S, et al. ,*Feasibility and efficacy of patient-initiated mobile teledermoscope for short-term monitoring of clinically atypical nevi.* JAMA Dermatol 2015;151(5):489–96.

[11] https://guides.lib.unc.edu/systematic-reviews/library-help.

[12] https://libguides.murdoch.edu.au/systematic/PICO.

[13] Amira S Ashour, Yanhui Guo, Enver Kucukkulahli, Pakize Erdogmus & Kemal Polat 2018, ,*‘A hybrid dermoscopy images segmentation approach based on neutrosophic clustering and histogram estimation’.* Applied Soft Computing, vol. 69, pp. 426-434.

[14] SeetharaniMurugaiyan Jaisakthi, Palaniappan Mirunalini & Chandrabose Aravindan 2018,*‘Automated Skin Lesion Segmentation of Dermoscopic Images using GrabCut and k-means algorithms’* The Institution of Engineering and Technology, doi: 10.l049/iet- cvi.2018.5289

[15] Sahar Sabbaghi M, Mohammad Aldeen, Senior Member, William V. Stoecker, Rahil Garnavi 2018,,*‘Biologically Inspired QuadTreeColourDetectioninDermoscopyImagesofMelanoma*, IEEEJournal of Biomedical and Health Informatics, doi: 10.1109/JBHI. 2018.2841428

[16] Brammya, G,Praveena, S,NinuPreetha, NS,Ramya, R,Rajakumar BR & Binu,D2018,,*‘DeerHuntingOptimizationAlgorithm:ANew Nature-Inspired Meta-heuristic Paradigm’*, Section A: Computer Science Theory, Methods and Tools, doi: 10.1093/comjnl/bxyl33

[17] Amira Soudani & Walid Barhoumi 2019,,*‘An Image-Based Segmentation Recommender using Crowdsourcing and Transfer Learning for Lesion Extraction’*, Expert Systems with Application, vol. 118, pp.400-410, doi:https://doi.org/10.1016/j.eswa.2018.10.029

[18] Walker,BN,Rehn,JM,Kalra,A,Winters,RM,Drews,P,Dascalu,I, David, EO & Dascalu, A 2019,,*‘DermoscopyDiagnosis of Cancerous Lesions Utilizing Dual Deep Learning Algorithms via Visual and Audio (Sonification) Outputs:* Laboratory and Prospective ObservationalStudies’,EBioMedicine,doi:https://doi.org/10.1016/ j.ebiom.2019.01.028

[19] Teck Yan Tan, LiZhang & Ming Jiang2016,,*‘An Intelligent Decision Support SystemforSkinCancerDetectionfromDermoscopic Images’, International Conference on Natural Computation.*, Fuzzy Systems and Knowledge Discovery,pp. 2194-2199, ISSN: 978-1-5090-4093-3/16.

[20] Mohammed,A,Al-Masni,Mugahed,A,Al-Antari,Mun-TaekChoiand Seung Moo-Han 2018,,*‘Skin Lesion Segmentation inDermoscopy Images via Deep Full Resolution Convolutional Networks’*, Computer Methods and Programs in Biomedicine, vol. 162, pp. 221-231, doi: 10.1016/j.cmpb.2018.05.027

[21] AnujKumar &UmeshChandra 2018,,*‘Comparative AnalysisofImageSegmentation using Edge-Region Based Technique and Watershed Transform’* International Journal ofLatest Technology inEngineering, Management & Applied Science, vol. 7,no. 5,ISSN: 2278-2540.

[22] GuotaiWang, WenqiLi, MariaAZuluaga, RosalindPratt, PremalAPatel, MichaelAertsen, TomDoel, AnnaLDavid, JanDeprest, Sebastien Ourselin & Tom Vercauteren 2018,*‘Interactive Medical ImageSegmentationusingDeepLearningwithImage-specificFine- tuning’*, IEEE Transactions on Medical Imaging, doi: 10.1109/TMI.2018.2791721.

[23] Qaisar Abbas, Irene Fondon Garcia, Emre Celebi, M & Waqar Ahmad 2011,,*‘A Feature-Preserving Hair Removal Algorithm for Dermos copy Images’*, Skin Research and Technology,Vol.0, pp.1-10, doi: 10.1111/j.1600-0846.2011.00603.x.

[24] Euijoon Attn, Jinman Kim, Lei Bi, Ashnil Kumar, Changyang Li, MichaelFulham & DavidDaganFeng2017,,*‘Saliency-basedLesion Segmentation via Background Detection in Dermoscopic Images’*, IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2017.2653179

[25] Roberta B Oliveira, Norian Marranghello, Aledir S Pereira, João Manuel & Tavares, RS 2016,,*‘A computational approach for detecting pigmented skin lesions in macroscopic images’*, vol. 61, pp. 53-63

[26] Eliezer Flores & Jacob Scharcanski 2016,,*‘Segmentation of melanocytic skin lesions using feature learning and dictionaries?’*, Expert Systems with Applications, vol. 56, pp. 300-309.

[27] Andrea Sboner, Claudio Eccher, Enrico Blanzieri, Paolo Bauer, Mario Cristofolini, Giuseppe Zumiani & Stefano Forti 2003,,*‘A multiple classifier system for early melanoma diagnosis’*, Artificial Intelligence in Medicine, vol. 27, no. 1, pp. 29-44.

[28] Leonardo Rundo, Alessandro Stefano, Carmelo Militello, Giorgio Russo, Maria Gabriell, Sabini, Corrad, D’Arrigo, Francesco Marletta, Massimo Ippolito, Giancarlo Mauri, Salvatore Vitabile & Maria Carl Gilardi 2017,,*‘A fully automatic approach for multimodal PET and MR image segmentation in gamma knife treatment planning’*, Computer Methods and Programs in Biomedicine, vol. 144, pp. 77-96

[29] Haidi Fana, FengyingXie,Yang Li, Zhiguo Jiang & Jie Liu 2017,,*‘Automatic segmentation of dermoscopy images using saliency combined with Otsu threshold’*, Computers in Biology and Medicine, vol. 85, pp. 75-85

[30] Mohammed A Al-masni, Mugahed A Al-antari, Mun-Taek Choi, Seung-MooHan & Tae-Seong Kim 2018, ,*‘Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks’*, Computer Methods and Programs in Biomedicine, vol. 162, pp. 221-231

[31] Neda Zamani Tajeddin & Babak Mohammadzadeh Asl 2018,,*‘Melanoma recognition in dermoscopy images using lesion’s peripheral region information’*, Computer Methods and Programs in Biomedicine, vol. 163, pp. 143-153.

[32] Manjunath Rao, Calvin Joshua Fernandez & Sreekumar, K 2020,,*‘Analysis of Melanoma Lesion Images using Feature Extraction & ClassificationAlgorithms’*, International JournalofRecentTechnologyandEngineering,vol.8,no.6,ISSN:2277-3878.

[33] Vikash Yadav & Vandana Dixit Kaushik 2018,,*‘Detection of MelanomaSkinDiseasebyExtractingHighLevelFeaturesforSkin Lesions’*,InternationalJournalofAdvancedIntelligenceParadigms, vol.11,No.’/4,pp.397-408.

[34] Kimia Kiani & Ahamad R Sharafat 2011, ,*'E-Shaver: An Improved DullRazorforDigitallyRemovingDark-andLightColoredHairsinDermoscopic Images'*,ComputersinBiologyandMedicine, Vo1.41, pp. 139-145, doi: 10.1016/j.compbiomed.2011.01.003.

[35] SinaKhakabi,Pau1Wightona, TimKLee & StellaAtkins,M2012,,*'Multi-level Feature Extraction for Skin Lesion Segmentation in Dermoscopic Images'*, Medical Imaging 2012: Computer Aided Diagnosis,doi:10.1117/12.911664

[36] Omkar ShridharMurumkar& Gumaste P. P 2015,,*'Feature Extraction forSkinCancerLesionDetection'*,International JournalofScience, EngineeringandTechnology Research,vo1.4,no.5,ISSN:2278— 7798

[37] Sharmin Majumder & Muhammad Ahsan Ullah 2018,,*'Feature Extraction from Dermoscopy Images for an Effective Diagnosis of MelanomaSkinCancer'*,InternationalConferenceonElectricalandComputer Engineering, ISSN: 978-1-5386-7482-6/18

[38] Xie, F.; Fan, H.; Li, Y.; Jiang, Z.; Meng, R.; Bovik, A. Melanoma,*Classification on Dermoscopy Images Using a Neural Network Ensemble Model.* IEEE Trans. Med. Imaging 2017, 36, 849–858. [CrossRef]

[39] Masood, A.; Al-Jumaily, A.A.; Adnan, T.,*Development of Automated Diagnostic System for Skin Cancer: Performance Analysis of Neural Network Learning Algorithms for Classification.* In Artificial Neural Networks and Machine Learning–ICANN 2014

[40] Cueva,W.F.; Munoz, F.; Vasquez, G.; Delgado, G.,*Detection of Skin Cancer "Melanoma" through Computer Vision.* In Proceedings of the 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), Cusco, Peru, 15–18 August 2017; pp. 1–4. [CrossRef]

[41] Jaleel, J.A.; Salim, S.; Aswin, R. Artificial Neural Network Based Detection of Skin Cancer. Int. J. Adv. Res. Electr. Electron. Instrum. Eng. 2012, 1, 200–205.

[42] Choudhari, S.; Biday, ,*S. Artificial Neural Network for SkinCancer Detection.* IJETTCS 2014, 3, 147–153.

[43] Aswin, R.B.; Jaleel, J.A.; Salim, S.,*Hybrid Genetic Algorithm: Artificial Neural Network Classifier for Skin Cancer Detection.* In Proceedings of the 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kanyakumari, India, 10–11 July 2014; pp. 1304–1309. [CrossRef]

[44] Feng-Ying, Xiea Shi-Yin Qin, Zhi-Guo Jiang & Ru-Song Meng 2009,,*'PDE-based unsupervised repair of hair-occluded information in dermoscopy images of melanoma'*, Computerized Medical Imaging and Graphics, vol. 33, no. 4, pp. 275-282.

[45] Suleiman Mustafa and Akio Kimura,,*'A SVM-based diagnosis of melanoma using only useful image features'*, 2018 International Workshop on Advanced Image Technology (IWAIT)

[46] S. R. Komatireddy, K. Meghana, V. Gude and G. Ramesh,,*"Facial Shape Analysis and Accessory Recommendation: A Human-Centric AI Approach,"* 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2023, pp. 182-191, doi: 10.1109/ICIMIA60377.2023.10426487.

[47] Venkataramaiah Gude, Sujeeth T, K Sree Divya, P. Dileep Kumar Reddy, G. Ramesh. (2024).,*Machine Learning for Characterization and Analysis of Microstructure and Spectral Data of Materials.* International Journal of Intelligent Systems and Applications in Engineering, 12(21s), 820–826.

[48] Hongming Xu, Cheng Lu, Richard Berendt, Naresh Jha & Mrinal Mandal 2018, ,*'Automated analysis and classification of melanocytic tumor on skin whole slide images'*, Computerized Medical Imaging and Graphics, vol. 66, pp. 124-134.

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Mar 12, 2024
Accepted: May 27, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 64–83, DOI 10.12694/scpe.v26i1.3529

DESIGNING AN INTUITIVE HUMAN-MACHINE INTERFACE FOR A SKIN CANCER DIAGNOSTIC SYSTEM: AN ENSEMBLE LEARNING APPROACH

PRASANNA LAKSHMI AKELLA[*]AND R KUMAR[†]

Abstract. In the domain of medical diagnostics, the efficacy of Human-Machine Interfaces (HMIs) plays a pivotal role in harnessing advanced computational models for practical clinical application. This study introduces a refined Ensemble Learning-Based Decision Support System, designed with an emphasis on intuitive HMI for accurate melanocytic and non-melanocytic skin cancer diagnosis. We present "EffiViT," a model that synergizes EfficientNet's robust feature extraction capabilities with the Vision Transformer's attention-based contextual understanding, tailored through an interface that prioritizes ease of use and interpretability for medical professionals. Through extensive evaluation on the ISIC 2019 benchmark dataset, EffiViT demonstrated a classification accuracy of 99.4%, coupled with superior performance in specificity and area under the ROC curve. The system's interface design was iteratively refined based on feedback from dermatologists, focusing on clear visualization of diagnostic information, straightforward navigation, and efficient access to model interpretations. Our findings underscore the importance of integrating user-centered design principles in the development of diagnostic tools, highlighting how a well-conceived HMI can enhance the adoption and effectiveness of AI-based systems in clinical settings. The proposed system stands out not only for its diagnostic accuracy but also for its contribution to the realm of HMI, offering insights into designing interfaces that facilitate better decision-making and ultimately improve patient outcomes in the field of dermatology.

Key words: Skin cancer, Ensemble model, Feature Extraction, Vision Transformer, Data Augmentation, Diagnostic Accuracy, Medical Imaging

1. Introduction and examples. Skin cancer poses a significant global health concern characterized by abnormal skin cell growth and the frequent emergence of malignant tumors. It primarily manifests in areas exposed to sunlight, often linked to ultraviolet (UV) radiation. Although prevalent in sun-exposed regions, it can also occur in areas with limited sunlight. The three primary types of skin cancer are basal cell carcinoma, squamous cell carcinoma, and melanoma, the latter being the most serious and highly dangerous. The alarming frequency of one skin cancer diagnosis every 57 seconds underscores the need for improved screening techniques and the integration of Computer-Aided Diagnosis (CAD) systems into clinical workflows [1].

In the United States, daily reports indicate over 9,500 new cases of skin cancer [2]. Projections by the American Cancer Society estimate 97,610 new cases of melanoma in 2023, with 7,990 fatalities [3]. Individuals with fair skin, light features, a history of sunburns, or a family history of skin cancer face higher risks. Additional risk factors include compromised immune systems, exposure to chemicals, and radiation therapy. These statistics and risk factors underscore the urgent need for innovative detection and diagnosis techniques, especially for high-risk groups.

Early detection is crucial for minimizing scarring and disfigurement associated with various forms of skin cancer. Therefore, effective management involves techniques for early detection and prevention. Recent advancements in medical image processing, particularly in the identification and classification of skin lesions, have significantly improved diagnostic accuracy [4]. CAD systems address challenges in skin lesion inspection, considering lesion localization, and the presence of hair. These systems aim to assist medical professionals in early detection, automatic identification of malignant lesions, and more efficient treatment.

This study proposes and evaluates a specific Ensemble model to enhance early detection, thereby improving patient outcomes. Recent advances in medical image processing, driven by machine learning and deep learning, have shown remarkable success in diagnosing diseases such as COVID-19 and pneumonia [5, 6]. Both supervised

[*]Department of Electronics and Instrumentation Engineering, National Institute of Technology Nagaland, Dimapur-797103,Nagaland, India, Mail ID: (prasannalakshmi.akella@gmail.com)

[†]Department of Electronics and Instrumentation Engineering, National Institute of Technology Nagaland, Dimapur-797103, Nagaland, India

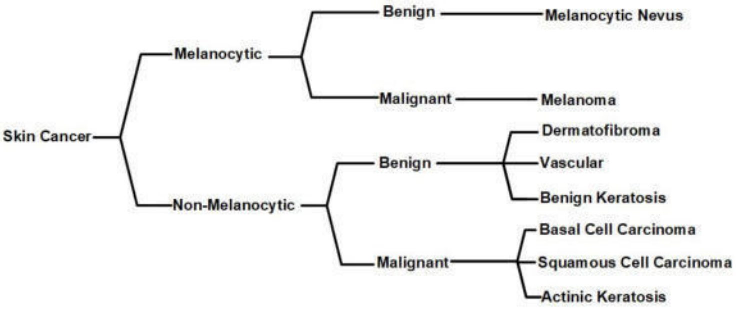


Fig. 1.1: Comprehensive Overview of ISIC 2019 Skin Cancer Types.

and unsupervised methods have been employed in deep learning models for detecting lung and pancreatic cancers [7]. Technologies like AlexNet [8] and VGG16 [9] have demonstrated impressive performance in various applications, including the identification of pulmonary diseases, facial recognition, and unmanned aerial vehicle photography. The various types of skin cancer, each with its unique characteristics, are considered in this study. The proposed Ensemble model, EfficientNetV2 B0-ViT, leverages deep learning technologies to revolutionize the detection and classification of skin cancer, ultimately enhancing patient care and treatment planning. Figure 1.1 illustrates the various skin cancer types examined in the study.

2. Major Contributions and Manuscript Organization. The significant contributions of the EffiViT Ensemble model in the realm of multiclass skin cancer detection are summarized as follows:

- Introduction of the EffiViT Ensemble model, a novel approach combining the strengths of EfficientNet and Vision Transformer (ViT) to advance multiclass skin cancer classification.
- Application of diverse image data augmentation techniques, including rotation, flip, zoom, and noise addition, to enhance the model's classification precision by expanding the dataset.
- Comprehensive evaluation of the model's performance using the ISIC 2019 benchmark dataset, enabling precise comparisons with established methodologies.
- Demonstration of the model's high accuracy in classifying various forms of skin cancer, underscoring its effectiveness for diagnosis and aiding in treatment planning decisions.

The subsequent sections of the paper are organized as follows: Section 3 presents a comprehensive Literature Survey, which critically reviews existing research and advancements in skin cancer classification methodologies. The Materials and Methods, detailed in Section 4, are subdivided into several parts, discussing dataset preparation, preprocessing techniques, data augmentation, and the specifics of the proposed EffiViT Ensemble model that combines EfficientNet and Vision Transformer architectures. Section 5, Results and Discussion, presents a robust evaluation of the model, encompassing experimental setup, performance metrics, confusion matrix analysis, classification reports, ROC-AUC curve analysis, and a comparative analysis with state-of-the-art models. Finally, the manuscript concludes with Section 6, summarizing the study's findings, highlighting the potential of the EfficientNet-ViT Ensemble model in enhancing diagnostic accuracy for skin cancer, and suggesting avenues for future research.

3. Literature Survey. The World Health Organization anticipates that by 2030, cancer will become the predominant cause of death, accounting for an estimated 13.1 million fatalities [10]. Recognizing the global scale of this challenge, research in skin cancer classification has taken on an international dimension, with significant contributions emerging from diverse regions. These studies address the classification across a spectrum of skin types and ethnic groups, underscoring the need for versatile and inclusive diagnostic solutions. Skin cancer is particularly prevalent, arising from abnormal cell proliferation that can swiftly invade and spread throughout the body [11].

A variation of methods for skin cancer classification have been devised and executed in the healthcare field in recent years. For instance, using the HAM10000 dataset, Chowdhury et al. [12] developed a CNN model

Table 3.1: Literature Survey Summary

Study	**Method**	**Dataset Used**	**Classes**	**Accuracy**
Chowdhury et al. [12]	Custom CNN	HAM10000	7	82.7%
Esteva et al. [13]	CNN	ISIC 2018	7	N/A
Li et al. [14]	VGG16 and ResNet-50	ISIC 2018	7	85%
Nunnari et al. [15]	VGG16 and ResNet-50	ISIC 2019	8	72.2%
Sadeghi et al. [16]	ResNet-50	N/A	4	60.94%
Xie et al. [17]	Deep CNN	ISIC 2017 and PH2	3	90.4%
Yang et al. [18]	ResNet-50	ISIC 2017	2	83%
Zunair et al. [19]	VGG 16	ISIC 2016	2	N/A
Kassem et al. [20]	Google Net	ISIC 2019	8	94.92%
Kasani et al. [21]	Transfer Learning	ISIC 2019	8	92%
Salido et al. [22]	CNN	PH2	2	93%
Shahin et al. [23]	Inception V3 and ResNet-50	ISIC 2018	N/A	89.9%
Sherif et al. [24]	Deep CNN	ISIC 2018	N/A	96.67%
Unver et al. [25]	YOLO and Grab Cut	PH2 and ISBI 2017	N/A	93.39%

that can identify seven kinds of skin diseases. Overall, their technique was 82.7% accurate and 78% precise. In their study, Esteva et al. [13] deployed a Convolutional Neural Network (CNN) to effectively discern seven distinct classes from the ISIC 2018 dataset. Their model achieved an impressive Area Under the Curve (AUC) metric of 94%, indicating its robust performance in accurately classifying the given data. Similarly, Li et al. [14] employed an Ensemble model of ResNet-50 and VGG16 to classify seven skin disease classifications with an accuracy of 85% using the ISIC 2018 dataset.

Nunnari et al.[15] used the ISIC 2019 dataset to classify eight different types of skin. Their rate of accuracy for the explanatory models, VGG16 and ResNet-50, were 72.2% and 76.7%, respectively. Using ResNet-50, Chilana et al. [16] successfully classified 1021 dermoscopy pictures into four skin types with an accuracy of 60.94%. Using a tweaked deep CNN, Xie et al. [17] successfully diagnosed three skin illnesses on the ISIC 2017 and PH2 datasets with an average accuracy of 90.4%. In order to classify two skin illnesses from the ISIC 2017 dataset, Yang et al. [18] employed ResNet-50 and achieved an accuracy of 83%. On the ISIC 2016 dataset, two skin conditions were classified using VGG16 by Zunair et al. [19]. An area under the curve of 81.18% and a sensitivity of 91.76% were obtained.

Kassem et al.[20] on the ISIC 2019 dataset to classify skin lesions into eight categories, proving that image augmentation and transfer learning improve classification accuracy. They achieved accuracy of 94.2%, precision of 73.62% sensitivity of 96.5%, Specificity of 73.62% and F1 Score of 74.04%. After more image enhancement and tweaks to the Google Net's architecture, the accuracy, sensitivity, specificity, precision, and F1 score improved to 94.92%, 79.8%, 97%, 80.36%, and 80.07%, respectively. In order to test several deep learning architectures for melanoma diagnosis, Kasani et al. [21] used image pre-processing to improve image quality and remove noise. To avoid overfitting, they added more data and found that the classification outcomes were considerably improved. 92% precision, 92% recall, and 93% accuracy were all achieved. Classifying skin lesions autonomously was established by Salido et al.[22] A deep CNN improved classification accuracy by 93% and 84% sensitivity by removing noise and artefacts from the image. Shahin et al. [23]used the Inception V3 and ResNet-50 architectures to make a system for classifying skin lesions based on deep neural networks. They trained on the ISIC 2018 dataset and reached validation accuracy rates of up to 89.9 percent. On the ISIC 2018 dataset, the accuracy of the deep CNN used by Sherif et al. [24] to classify and detect melanomas was 96.67%.

For melanoma identification, Unver et al. [25] used the most recent deep learning system. You Only Look Once (YOLO) was utilized for detection, and then Grab Cut was used for cutting out unwanted parts. Using the PH2 and ISBI 2017 datasets, they were able to achieve a precision of 93.39 %.

A summary of literature studies is presented in the table 3.1.

Current study shows a great variety of skin cancer classification methods, emphasizing the importance of machine learning, particularly deep learning. Skin cancer classification accuracy has been improved using

Convolutional Neural Networks (CNNs), VGG16, ResNet-50, and different Ensemble models. The literature reviews observation on methodological variation between research is significant. It's important to highlight that these methodological discrepancies, especially in machine learning model selection and preprocessing, can affect skin cancer classification results. An in-depth discussion about these methodological variations might help explain why some procedures or models yielded better or worse results.

Skin cancer classification is continuously evolving, despite these advances. Current approaches can improve accuracy, precision, and memory, but they still need improvement. Most previous studies have focused on binary classification; however, classification can be expanded towards Multi class classification. In addition to highlighting these disparities, it is important to highlight the obstacles these studies face. Data collection, dataset biases, and overfitting or underfitting during model training can make skin cancer classification model creation and validation difficult. This paper proposes a novel Ensemble model that combines the strengths of EfficientNet and the Vision Transformer architectures to advance the field and overcome these gaps and obstacles. With its promising feature extraction and global dependency capture, this approach aims to improve classification performance. The Ensemble model contributes to skin cancer classification, and exploring other research avenues in the future may further enhance understanding. For instance, improving deep learning models, integrating them with medical diagnostic tools, or exploring new class imbalance methods could lead to beneficial research. In the following sections, the proposed model will be assessed, its efficacy examined, and contrasted with currently used field methods. Despite significant advances in the development of machine learning models for skin lesion classification, the current literature reveals critical limitations. These limitations underscore the need for robust model validation to enhance the accuracy and applicability of these models.

1. *Focus on Binary Classification:* Many studies predominantly focus on binary classification. This approach may not adequately capture the complexities involved in the multiclass categorization of skin lesions, which is necessary to address the diverse nature of skin diseases.
2. *Variations in Methodological Approaches:* There is considerable variability in model selection and data preprocessing across studies. These variations can significantly impact the consistency and accuracy of classification outcomes, leading to potentially unreliable results.
3. *Data Collection and Dataset Biases:* Issues related to data collection and inherent biases in datasets pose significant challenges. These biases affect the generalizability and applicability of the models, making them less effective in real-world scenarios.
4. *Risks of Overfitting or Underfitting:* Many models face risks of overfitting or underfitting during the training phase. This highlights the importance of developing more adaptable and resilient machine learning models that can perform well across various conditions.

4. Materials and Methods.

4.1. Acquiring Dermoscopic Images of Skin Lesions. In this paper, the effectiveness of the proposed method for skin cancer classification is evaluated using the publicly available dataset ISIC 2019. This dataset consists of 25,331 RGB images that offer a comprehensive set of cases for evaluation, spanning eight classes: melanocytic nevus (NV), melanoma (MEL), dermatofibroma (DF), vascular lesion (VASC), benign keratosis (BKL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and actinic keratosis (AKIEC). These classes encompass a wide range of skin cancer types, making the dataset an excellent resource for training robust classification models. Table 4.1 presents the class distribution within the ISIC 2019 dataset, including the number of samples for each class. While the ISIC 2019 dataset serves as an excellent resource for this study, its comprehensive array of images and annotations, which span a wide spectrum of skin cancer types, makes it an ideal choice. This dataset not only offers a robust training environment but also ensures that our model is tested against a diverse set of diagnostic scenarios, enhancing its ability to generalize well across different skin cancer classes. However, reliance on this single dataset may limit exposure to variations found in broader clinical settings.

Before training the classification model, the images were standardized to ensure uniform input data quality. This involved resizing all images to 224x224 pixels using bilinear interpolation, normalizing the pixel values, and performing color space conversions when necessary. Additionally, data augmentation techniques, such as random rotations, flips, and zooms, were applied to increase the diversity and variability of the training data. By using augmentation approaches, the problem of overfitting is avoided, improving the model's ability

Table 4.1: ISIC 2019 Dataset Class Distribution

Class & Abbreviation	Number of Samples
Melanocytic Nevus (NV)	12,875
Melanoma (MEL)	4,522
Dermatofibroma (DF)	239
Vascular Lesion (VASC)	253
Benign Keratosis (BKL)	2,624
Basal Cell Carcinoma (BCC)	3,323
Squamous Cell Carcinoma (SCC)	628
Actinic Keratosis (AKIEC)	867

to generalize to unseen data. By leveraging the preprocessed and augmented datasets, separate training and testing were conducted on each dataset. The experimental setup involved allocating 80% of the images for training and 20% for validation and testing. This distribution ensures a comprehensive learning process while providing a robust means to evaluate the model's predictive capabilities.

4.2. Data pre-processing. In the framework, thorough data preparation procedures were implemented to assure the best possible state of the multi-class skin cancer dataset for classification accuracy. To commence, each image underwent a resizing process to a uniform resolution of 224 x 224 pixels using bilinear interpolation. This specific size was chosen to balance detail preservation with computational efficiency, making it well-suited for the deep learning models used. The images were resized using bilinear interpolation, which improved their quality and maintained their visual integrity. This step was crucial in preserving the diagnostic features of the skin lesions. To further refine the images, the median filtering technique was utilized to eradicate any noise present in the data.

Median filtering was specifically applied to reduce salt-and-pepper noise, which is common in dermatological images due to variations in lighting and camera quality. Normalizing the pixel values with min-max scaling ensures consistency and comparability across the dataset. Normalization was applied to all channels of the images, adjusting the pixel values to a standard scale that enhances the algorithm's sensitivity to subtle variations in skin lesions. Finally, hair artifacts were removed from the images using a filtering technique known as Blackhat filtering, which effectively obliterated any unwanted hair-like structures. The Blackhat filtering was complemented by a custom algorithm designed to detect and subtract complex hair patterns without affecting underlying lesion details, ensuring that the diagnostic features remain unobscured.

These preprocessing steps optimized the images for subsequent analysis and interpretation by contributing to their enhancement and refinement. Utilizing bilinear interpolation, each image within the dataset was scaled uniformly to a resolution of 224 * 224 pixels. To achieve a balance between the preservation of essential details and the limitation of computational resources, this particular scaling technique was employed. Median filtering was utilized to reduce the presence of objectionable artifacts and noise, thereby improving the image's overall quality. Through the effective reduction of stochastic noise, the application of this technique has resulted in an appreciable improvement in the clarity and coherence of photographs. The goal was to improve the model's ability to extract pertinent data for precise categorization. The pixel values were then normalized using min-max scaling, which effectively rescaled them to suit within the range [0, 1]. This normalization technique has facilitated the accomplishment of consistent training outcomes by standardizing pixel values and fostering convergence across a multitude of features and channels. To reduce the possibility of hair artifacts interfering with accurate classification, a concerted effort was made to eradicate hair. Utilizing Blackhat filtering techniques, the hair filaments present in the epidermis photographs were effectively emphasized and then eliminated.

The primary objective of this phase was to refine the model's focus on the critical patterns associated with skin cancer by eliminating irrelevant data and characteristics. Figure 4.1 illustrates the data preprocessing workflow, showing sample images before and after the application of these techniques. This visual representation underscores the importance of preprocessing in improving image quality for accurate classification. The visual representations are displayed in the left column prior to preprocessing, and in the right column after resizing,

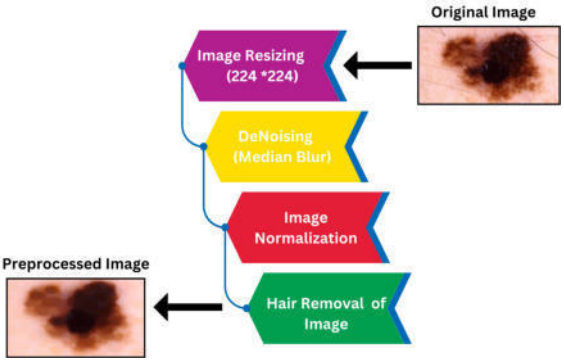


Fig. 4.1: Data Preprocessing Workflow for Skin Cancer Classification. This diagram illustrates the steps involved in preparing data for the analysis and classification of skin cancer, highlighting the crucial preprocessing stages required for effective deep learning model training.

noise reduction, and hair removal techniques have been applied. The provided visual examples demonstrate the effectiveness of employing preprocessing techniques to improve image quality and reduce objectionable artifacts such as hair distortions, among other notable benefits.

The impact of these data preprocessing steps on the subsequent skin cancer classification model's performance was evaluated during the training and evaluation phase. The comprehensive preprocessing workflow aimed to improve model accuracy, robustness, and generalization by minimizing noise, standardizing features, and eliminating hair artifacts. Performing data preprocessing techniques, including resizing, noise removal, normalization, and hair removal, ensured the dataset's quality and suitability for effective multi-class skin cancer classification.

4.3. Data Augmentation. The skin cancer classification model was improved using skin cancer image data augmentation techniques. Data augmentation reduces training data and skin lesion appearance issues. Many augmentation methods were employed to increase dataset diversity and unpredictability. Table 4.2 presents a summary of the original dataset and the augmentation process. Significantly, augmenting the dataset resulted in a substantial increase in the number of images, a pivotal factor contributing to the robustness of the classification model. Random rotations between -10 and 10 degrees and horizontal and vertical flips were applied to create skin cancer photo orientations and mirror image variations. Tiny translations and random zooming simulated scale and viewpoint. These changes improved the model's skin cancer classification. Domain-specific factors determined augmentation methods. Augmentation approaches were customized to highlight relevant changes that match the visual characteristics of different skin lesion types. Highlighting augmentation asymmetry or irregular edges helps train these qualities. Data augmentation categorizes skin tumors empirically. Adding different variants to the dataset has been found to address the problem of insufficient data and increase the model's generalisation skills. Augmented data exposes the model to more skin lesion appearances, helping it acquire robust and discriminative skin cancer features.

Figure 4.2 shows data augmentation-enabled changes in the supplemented dataset. The left column shows original skin lesions, whereas the right column shows augmented ones. These examples show how augmentation tactics create a richer dataset. This part describes the augmentation methods, including translation, zooming, and rotation parameter adjustments. Each augmentation method's explanation and its impact on the model's ability to capture various visual features of skin cancer are explored. Data augmentation was utilized to enhance the skin cancer classification model, aiming to efficiently handle skin lesion appearances and improve classification accuracy.

4.4. Proposed Methodology. The methodology proposed in this study is for the diagnosis of skin cancer entails the application of a multiclass dataset. In order to tackle the difficulties arising from a scarcity of training data and the inherent variability in the visual characteristics of skin lesions, data augmentation techniques are

Table 4.2: Summary of Data Augmentation

Lesion Type	Images Before	Augmentation	Augmented Images
NV	12,875	NO	12,875
MEL	4,522	NO	4,522
DF	239	YES	3,476
VASC	253	YES	4,281
BKL	2,624	NO	2,624
BCC	3,323	NO	3,323
SCC	628	YES	3,423
AK	867	YES	3,476

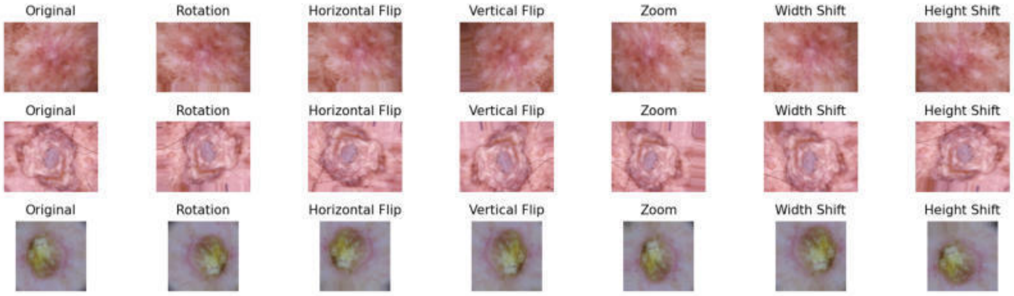


Fig. 4.2: Sample Augmented images

employed. Following the preprocessing stage, two image classification models, namely EfficientNet V2 B0 and ViT-B16, are trained utilizing transfer learning techniques. The augmented dataset was employed to introduce a diverse range of skin lesion presentations, thereby improving the robustness of the classification process. The training procedure includes selecting a loss function with symmetric cross-entropy [27]. Additionally, the Rectified Adam optimizer, which is recognized for its enhanced convergence and efficiency, is selected for optimizing the model.

The Geometric Mean Ensembling technique was utilized to combine the two models into an Ensemble framework, optimizing the strengths of both architectures for superior classification performance. The proposed approach involves assigning suitable weights to the predictions of each model, taking into account their respective performance on the validation set. This strategy aims to capitalize on the unique capabilities of both the EfficientNet and ViT models. The utilization of this Ensemble model allows medical practitioners to leverage computer-aided diagnosis, thereby augmenting the precision and dependability of skin cancer diagnosis. The performance of the model was assessed and measured using a range of metrics, such as Accuracy Score, Precision, Recall, and F1-score, on an independent validation/test dataset. The schematic representation of the entire procedure, encompassing data augmentation, training, and the construction of an Ensemble model, is depicted in Figure 4.3. The proposed methodology effectively encompasses the diverse visual attributes exhibited by skin cancer lesions, thereby augmenting the model's efficacy through the inclusion of a more diverse dataset.

4.5. Models Description of the Proposed Methodology. This section presents a comprehensive description of the models employed in the suggested approach for detecting skin cancer. The Ensemble model synergizes the capabilities of EfficientNet and Vision Transformer (ViT) architectures, aiming to leverage the strengths of each to enhance diagnostic performance.

4.5.1. EfficientNet V2 B0. EfficientNet V2 B0 [26] represents a highly advanced CNN architecture meticulously designed to optimize the process of image classification, particularly in the domain of skin cancer categorization. The model employs Mobile Inverted Bottleneck Convolution (MBConv) and Fused Mobile

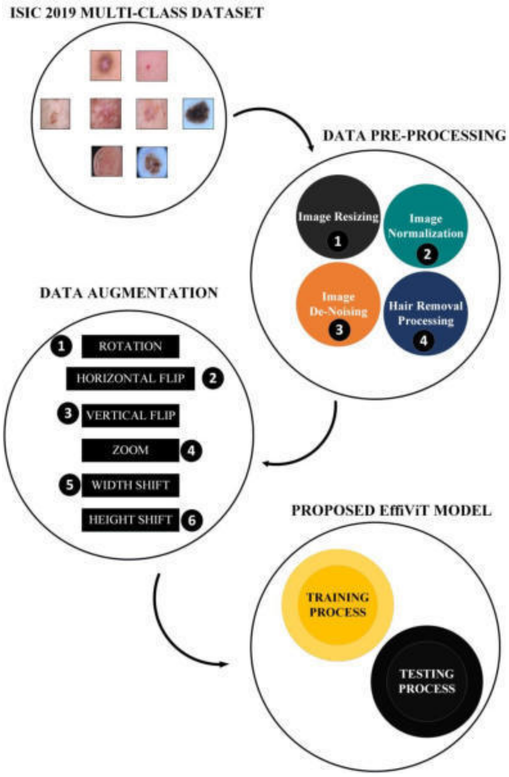


Fig. 4.3: Flow Graph of proposed methodology

Inverted Bottleneck Convolution (Fused-MBConv) blocks, enhancing efficiency and precision. Squeeze-and-Excitation (SE) blocks within these layers enable dynamic importance allocation to different channels, improving discriminative capabilities crucial for identifying skin cancer patterns. For this study, the Efficient Net V2 B0's adaptation involves fine-tuning with the ISIC 2019 dataset, optimizing it for high accuracy in skin cancer classification while maintaining computational efficiency.

Its primary objective is to enhance the efficiency and precision of such applications. EfficientNet V2 is a notable advancement that surpasses its predecessor, EfficientNet V1, by prioritizing the acceleration of training duration and enhancing the efficacy of parameters. The achievement of this objective is facilitated through the integration of compound scaling techniques with training-aware neural architecture search methodologies.

The architecture of EfficientNet V2 B0 incorporates crucial components such as the Mobile Inverted Bottleneck Convolution (MBConv) and Fused Mobile Inverted Bottleneck Convolution (Fused-MBConv) blocks. These elements play a significant role in the overall structure of the model. The MBConv blocks were derived from the MobileNetV2 inverted residual blocks. The architectural design of these blocks incorporates an expansion convolution layer subsequent to a depth-wise separable convolution layer. The Fused-MBConv blocks effectively optimize memory utilization and training duration by integrating the depth-wise and expansion convolutions into a unified standard 3x3 convolution block. These architectural components are indispensable for the extraction and manipulation of features from input images.

EfficientNet V2 B0 incorporates Squeeze-and-Excitation (SE) blocks within both the MBConv and Fused-MBConv layers, enabling the model to dynamically assign importance to different channels and enhance its discriminative capabilities. SE blocks utilize a mechanism called channel-wise relevance weights to dynamically recalibrate feature responses that are specific to each channel. Through the process of recalibration, the model is now able to focus its attention on the most pertinent features and discern crucial patterns linked to skin cancer lesions. EfficientNet V2 B0's design can be effectively elucidated through the use of a visual representation

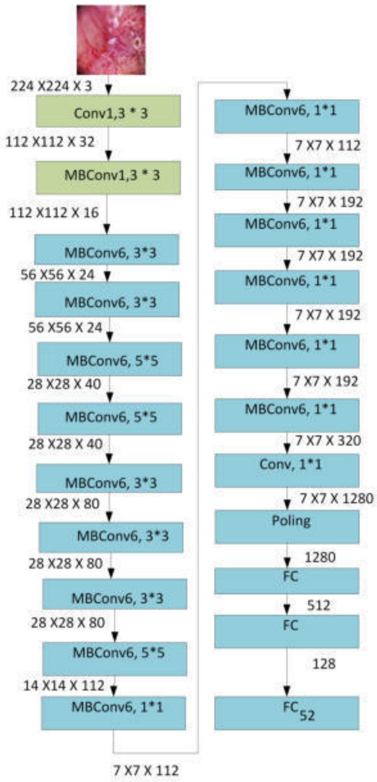


Fig. 4.4: EfficientNet V2 B0 Architecture with SE Blocks for Skin Cancer Classification

shown in Figure 4.4.

The structured visualization of the EfficientNet V2 B0 architecture, as depicted in Figure 5, highlights the sequential layers and data flow within the diagram. The input image is processed by standard convolutional blocks (Conv) in the initial layers. Subsequently, a sequence of MBConv layers is implemented, and these blocks are fundamental to the design of the network. The spatial dimensions and depth of the feature maps adjust in accordance with the data's progression through these layers, as depicted in the diagram. This modification is indicative of the network's compound scaling methodology. The depicted path from unprocessed image input to the ultimate classified output, in which the network completes its assignment of classifying skin cancer lesions, is visually represented in this roadmap. The incorporation of SE blocks into the network architecture, in conjunction with the MBConv layers, augments the model's emphasis on pertinent characteristics, a critical factor in ensuring precise skin cancer detection.

Compound scaling is a pivotal aspect of EfficientNet V2 B0, as it effectively governs the network's depth, width, and resolution. The subsequent information presents the equation for depth scaling.

$$\text{Number of layers in each block} = \text{round}(\alpha \cdot \beta^{\phi}) \tag{4.1}$$

In the equation for depth scaling in the EfficientNet V2 B0 model, each term plays a specific role in determining the architecture of the neural network. The coefficient α sets the baseline number of layers in the network. It acts as a hyperparameter, essentially determining the initial depth of the network. The parameter β is another crucial hyperparameter, typically utilized to control the scaling of the network's width, such as the number of channels in convolutional layers. The factor ϕ is used as a scaling factor, instrumental in adjusting the network's depth. Different values of phi correspond to different versions of the EfficientNet, each version varying in complexity and capacity.

The width scaling formula is given by:

$$\text{Number of channels in each block} = \text{round}(\gamma \cdot \alpha^{\phi}) \tag{4.2}$$

Where γ sets the baseline number of channels (or filters) in the convolutional layers of the network. Together, these equations embody the concept of compound scaling, a distinctive feature of EfficientNet, which ensures a balanced and efficient scaling of the network. This balanced approach optimizes the network's performance while maintaining computational efficiency, a key factor in the success and popularity of EfficientNet models.

4.5.2. Vision Transformer (ViT) B16 model. The ViT model represents a highly advanced methodology for the purpose of image classification tasks. It draws inspiration from the Transformer model, which has gained significant prominence in the field of natural language processing. In this study, ViT B16's application includes a focus on its self-attention mechanisms which are particularly effective in handling the detailed and varied patterns present in skin cancer images. Each encoder in the model uses multi-head attention and a feed-forward layer, enabling it to excel in image classification tasks by recognizing complex interdependencies. ViT B16 has been adapted to analyze skin cancer images, demonstrating its ability to efficiently process and classify large-scale dermatological data.

The ViT model, as opposed to traditional Convolutional Neural Networks (CNNs), leverages self-attention mechanisms to effectively capture extensive dependencies and overarching relationships within images. A multi-head attention layer and a feed-forward layer make up each encoder component. In the multi-head attention layer, attention weights are calculated by comparing how similar different image areas are. The aforementioned procedure can be formally expressed in mathematical notation as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.3}$$

In the Vision Transformer (ViT) B16 model, the self-attention mechanism is defined by the terms Q (Query), K (Key), and V (Value). The Query represents a specific part of the image, used to determine how much attention other parts of the image should receive. The Key assists in this process by comparing different parts of the image to the Query to calculate attention weights. The Value, representing the actual image content, is then scaled by these weights. This mechanism allows the ViT model to focus on the most relevant features within an image, capturing extensive dependencies and relationships. This approach, diverging from traditional CNNs, underscores the ViT's advanced capabilities in image classification tasks. By applying nonlinear transformations to the attention outputs, the feed-forward layer complements the multi-head attention layer. The formula that represents it is as follows:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{4.4}$$

In this equation, the input features are denoted by the variable x, and the learnable weight matrices and bias factors are denoted by W1, W2, b1, and b2. The input characteristics are altered by matrix multiplication with W1, then the bias term b1 is added. By introducing non-linearity, the ReLU activation function enables the model to recognize intricate patterns. The generated features then go through another matrix multiplication with W2 before being added to get the feed-forward layer's final output. By applying attention mechanisms and feed-forward layers, Figure 4.5 demonstrates how the ViT model performs significantly well on a range of image classification tasks.

Figure 4.5 represents the ViT B16 model in its initial stages, which involve the creation of image fragments from the input image. In order to preserve spatial information, which is essential for the model to comprehend the arrangement of the image, these regions are subjected to a linear projection in conjunction with positional embedding. In order to extract intricate patterns from the image data, the patches are subsequently processed by the Transformer encoder, which employs layers of multi-head self-attention and normalization. Before being fed into a SoftMax function for the final classification, the output from the encoder is normalized in batches, flattened, and passed through a dense layer.

The system demonstrates exceptional proficiency in capturing and analyzing the interconnections and inter-dependencies present within images. This enables it to effectively discern intricate patterns and accurately

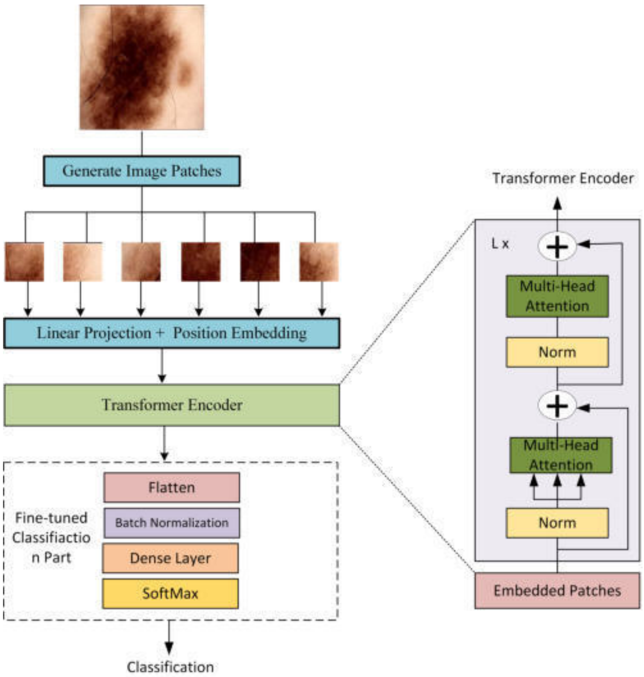


Fig. 4.5: Visualization of ViT Model Performance

classify images based on their inherent characteristics. The ViT model's remarkable scalability and adaptability position it as a preeminent methodology within the realm of computer vision. Its unique characteristics make it especially well-suited for effectively managing vast amounts of information on a large scale.

4.5.3. Ensemble via Averaging: Combining EfficientNet and ViT. In order to classify skin cancer using Ensemble methods, the EfficientNet and ViT Ensemble Model via averaging is a potent strategy to integrate the EfficientNet and ViT architectures. While the ViT excels in capturing comprehensive contextual information from global image regions, EfficientNet has gained significant recognition for its remarkable efficacy and scalability in various computer vision tasks. In this Ensemble model, multiple instances of the ViT and EfficientNet models are independently trained, and diversity is promoted by utilizing various initializations or hyperparameters for each model. During the inference process, the Ensemble generates predictions by employing a technique known as averaging, which involves computing the mean of the individual outcomes produced by each constituent model. This averaging process not only reduces variance among the model predictions but also maximizes the strengths of each architecture, ensuring a balanced approach to feature extraction and context analysis. The ensemble's predictions are further optimized through the utilization of a coefficient weighting approach. The ViT model employs a multiplication operation with a coefficient of 0.7 to scale its output probabilities, while the Efficient Net model utilizes a coefficient of 0.3 for the same purpose. These coefficients were meticulously calibrated based on extensive validation tests that measured the predictive efficacy of each model independently, ensuring that their contributions to the final decision are proportionate to their demonstrated reliability. The coefficients in the ultimate ensemble outcome indicate the proportional significance assigned to the predictions of each individual model.

The Ensemble model effectively consolidates the weighted outcomes to generate a conclusive prediction for the classification of skin cancer. The Ensemble methodology effectively addresses the challenge of reconciling the localized visual data obtained from EfficientNet with the broader global contextual information captured by ViT, primarily due to the implementation of a well-designed weighting strategy. The adjustment of coefficients is contingent upon the problem's inherent characteristics and the performance exhibited by individual models.

The ViT and EfficientNet models have successfully acquired a diverse range of representations through their training process. By leveraging these learned representations, the Ensemble model effectively enhances both the robustness and precision of skin cancer categorization. By implementing a systematic approach that focuses on mitigating the shortcomings of each model, this particular strategy effectively harnesses the unique capabilities and advantages possessed by each individual model.

5. Results and Discussion.

5.1. Experiment Environment. The hardware configuration for the experimental platform utilized in this paper is an Intel Xeon(R) CPU E5-2780 with a 2.80 GHz core frequency and an NVIDIA GeForce RTX 1080 GPU. Using the PyTorch framework [28], the suggested model is implemented in Python 3.7, ensuring a combination of high computational power and state-of-the-art software capabilities for handling deep learning tasks.

5.2. Evaluation Metrics. The evaluation of the classification model incorporated four metrics critical to medical diagnosis: Accuracy, Precision, Recall, and the F1-Score. The selection of these metrics was influenced by their importance in clinical decision-making. In this context, it's crucial not only to achieve high overall accuracy but also to effectively reduce the occurrence of both false negatives and false positives.

The accuracy of the model's predictions is the most fundamental performance metric. It can be defined as the ratio of accurately identified samples that are positive or negative to the total number of samples:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5.1}$$

Precision is a metric that focuses on the number of true positive predictions relative to the total number of positive predictions. It measures the model's ability to correctly identify positive instances:

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{5.2}$$

Recall is a metric that measures the ability of a model to correctly identify all positive instances:

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{5.3}$$

The F1 score is the harmonic mean of Precision and Recall:

$$\text{F1 Score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{5.4}$$

In the above formulas, TP represents true positive samples, TN for true negative samples, FP for false positive samples, and FN for false negative samples.

5.3. Evaluation of Proposed System. Firstly, the dataset is divided into three different sets: the training, Validation and the test sets. The division is performed in accordance with a predetermined ratio of 80:10:10. During training, a batch size of 32 and a learning rate of 0.0001 were employed to optimize the balance between comprehensive learning and computational efficiency. Both EfficientNet model and ViT model undergoes a training process that involves 50 epochs. The choice of 50 epochs was determined based on preliminary experiments that indicated this was an optimal balance between achieving sufficient model convergence and preventing overfitting, given the complexity of the models and the dataset size.

This approach was optimized to balance thorough learning and computational efficiency. After the completion of the training process, the model parameters were assessed by utilising the test dataset. The EffiViT model, which is an Ensemble model, demonstrated the successful integration of the EfficientNet and ViT models, using their respective strengths. The performance indicators acquired during the training and validation stages offer valuable insights into the model's ability to learn effectively and consistently throughout the learning process. As an example, during epoch 1, the training accuracy of EfficientNet-V2 B0 is recorded as 75.096%,

Table 5.1: Model Performance Summary

Epoch	EfficientNet - V2 B0				ViT- B16				EffiViT			
	Train Acc	Train Loss	Valid Acc	Valid Loss	Train Acc	Train Loss	Valid Acc	Valid Loss	Train Acc	Train Loss	Valid Acc	Valid Loss
1	75.096	1.0863	73.36	0.9125	31.29	1.3972	56.15	1.156	90.115	0.418	93.145	0.158
4	82.65	0.8543	79.46	0.8060	44.67	1.2012	62.83	1.032	92.00	0.350	94.500	0.125
8	84.716	0.7711	80.33	0.7540	57.82	0.9876	68.95	0.908	93.750	0.300	95.750	0.100
12	89.716	0.7149	82.66	0.7312	65.82	0.8542	72.34	0.794	95.250	0.250	96.500	0.085
17	90.145	0.6571	85.007	0.7180	71.18	0.7219	75.12	0.682	96.5500	0.200	97.000	0.070
20	91.635	0.6379	87.75	0.7012	74.56	0.6511	76.98	0.621	97.250	0.180	97.375	0.065
24	93.813	0.5508	90.145	0.6169	78.12	0.5834	79.24	0.562	97.750	0.160	97.625	0.060
28	94.847	0.5067	92.849	0.6010	80.89	0.5234	81.15	0.514	98.000	0.140	97.750	0.055
32	95.936	0.4903	93.183	0.5948	83.24	0.4821	82.72	0.479	98.250	0.120	97.875	0.050
37	96.118	0.4126	94.748	0.5585	85.47	0.4439	84.31	0.447	98.500	0.100	98.000	0.040
40	96.813	0.3893	94.999	0.4028	87.22	0.4123	85.64	0.419	98.750	0.080	98.055	0.030
44	97.145	0.3893	96.318	0.2855	88.75	0.3847	86.78	0.395	99.000	0.060	98.110	0.035
48	97.995	0.2545	97.113	0.2067	90.12	0.3585	87.82	0.372	99.100	0.040	98.165	0.030
50	98.9324	0.1976	97.491	0.1933	91.27	0.335	88.67	0.352	99.20	0.020	98.195	0.015

whereas the validation accuracy stands at 73.36%. In contrast, it can be observed that ViT-B16 demonstrates comparatively lesser accuracies, whereas EffiViT showcases the highest levels of accuracy. As the training advances to 50 epochs, a noticeable enhancement in accuracy and reduction in loss is observed across all models. Particularly, EffiViT consistently exhibits superior performance compared to the individual models.

By following these steps, the proposed method EffiVIT Ensemble model is used for skin cancer classification. This Ensemble model takes advantage of the strengths of both models to enhance classification performance and accuracy. Table 5.1 shows the performance metrics for the Ensemble Model, ViT-B16, and EfficientNet-V2 B0 models during training and validation.

Figure 5.1 illustrates the loss and accuracy curves for both the Vision Transformer model and the EfficientNet model throughout the training and validation phases. The graph on the left illustrates the Training and Validation Loss, wherein both losses exhibit a decreasing trend across the epochs, suggesting a notable enhancement in the model's performance. The EffiViT Ensemble model exhibits the lowest validation loss, indicating superior generalisation capabilities. The graph on the right displays the Training and Validation Accuracy. It is evident that the EffiViT Ensemble model exhibits the best accuracy, suggesting its superior predictive capabilities. The presented graphs highlight the superior learning capacity and effectiveness of the Ensemble methodology compared to the training of individual models.

In addition, the Ensemble model outperforms individual models in terms of inference time efficiency. While the EfficientNet V2 B0 model demonstrated a balance of speed and accuracy, the ViT- B16 excelled in capturing global dependencies, albeit at a slightly higher computational cost. The Ensemble model, through its strategic combination of both architectures, achieves an optimal balance of accuracy and inference time, suitable for real-time clinical applications The EfficientNet -V2 B0 model has an inference rate of 0.0059 seconds per sample and an average inference time of 1.66 seconds per sample, with a standard deviation of 0.93 seconds and 0.0033 seconds, respectively. The ViT-B16 model has an inference rate of 0.0089 seconds per sample and an average inference time of 2.49 seconds per sample, with a standard deviation of 0.08 seconds and 0.0003 seconds, respectively. With an inference rate of 0.0147 seconds per sample, the Ensemble Model, which combines the predictions of the two models, obtains an inference time of 4.15 seconds.

According to these inference time results, the Ensemble Model nevertheless maintains a fair inference rate, making it useful for real-world applications, even though the combining of numerous models makes it slightly more computationally time-consuming. The increased diagnostic performance and rapid inference time of the Ensemble Model demonstrate its potential as a reliable and efficient approach for diagnosing skin cancer.

5.4. Confusion Matrix Analysis. The confusion matrix, as shown in Figures 5.2, 5.3,5.4, offers a comprehensive view of the classification performance of the three models. While overall accuracy assessment is important, the detailed insights provided by the confusion matrix are crucial for understanding the classifica-

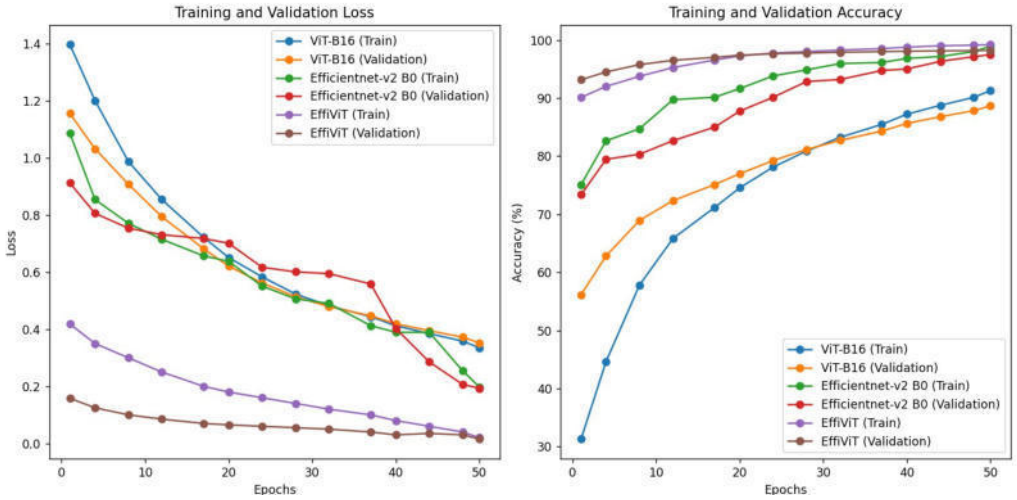


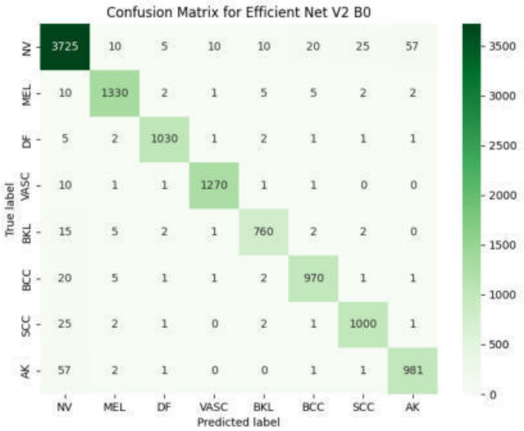
Fig. 5.1: Training and Validation Loss and Accuracy for Models

Table 5.2: Inference Time Efficiency Comparison

Model	Average Inference Time (s/sample)	Standard Deviation
EfficientNet-V2 B0	0.0059	0.0033
ViT-B16	0.0089	0.0003
Ensemble Model: EffiViT	0.0147	-

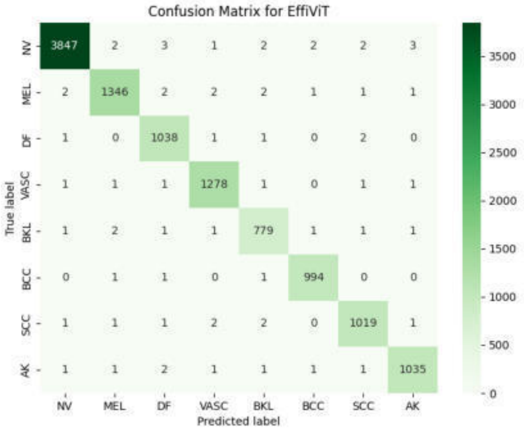
tion strengths and weaknesses of each model. The EfficientNet V2 B0 matrix(as seen in Figure 5.2 reveals its precision in classifying NV (Nevus) with 3725 true positives but also indicates a tendency to mistakenly categorize AK (Actinic Keratoses) as NV in 57 instances. Although Melanoma (MEL) and other categories like Dermatofibroma (DF), Vascular lesions (VASC), and Basal cell carcinoma (BCC) are generally well-identified, there are occasional errors, especially in differentiating Actinic keratosis (AK). The confusion matrix of ViT B16(as shown in Figure 5.3 has a continuous pattern characterised by a notable number of true positives for the NV class, along with equivalent performance for the MEL, DF, and VASC classes. Nevertheless, there is a marginal rise in misclassifications, particularly in distinguishing between AK and NV, which is a consistent pattern observed in all models.

The matrix of the EffiViT model(as shown in Figure 5.4 demonstrates notable enhancements, particularly in accurately classifying NV with 3847 instances correctly identified as true positives. Moreover, the model exhibits improved overall performance by reducing the number of misclassifications. It accurately recognizes MEL 1346 times, DF 1038 times, and other classes with great accuracy. It is noteworthy that the accuracy of AK exhibits a little enhancement, so highlighting the Ensemble model's aptitude for distinguishing among increasingly difficult categories. The utilization of a confusion matrix offers a comprehensive evaluation of the efficacy of various models in accurately categorizing distinct types of skin lesions. This analytical tool plays a crucial role in measuring the competency of these models in their classification tasks. The detailed information provided is of great value in formulating therapeutic strategies for the screening of skin cancer. The performance of the Ensemble Model is remarkable, exhibiting an accuracy rate of 99.4%. This surpasses the performance of existing models and signifies a noteworthy progression in the field of skin cancer classification. Furthermore, it has the potential to establish novel benchmarks in terms of diagnostic accuracy. The efficacy of the EffiViT Ensemble model is demonstrated by its capacity to minimize mis-classifications, specifically in distinguishing between AK and NV. This is achieved by leveraging the individual capabilities of the EfficientNet V2 B0 and ViT B16 models, resulting in a powerful combined model.



True label \ Predicted label	NV	MEL	DF	VASC	BKL	BCC	SCC	AK
NV	3725	10	5	10	10	20	25	57
MEL	10	1330	2	1	5	5	2	2
DF	5	2	1030	1	2	1	1	1
VASC	10	1	1	1270	1	1	0	0
BKL	15	5	2	1	760	2	2	0
BCC	20	5	1	1	2	970	1	1
SCC	25	2	1	0	2	1	1000	1
AK	57	2	1	0	0	1	1	981

Fig. 5.2: Model Comparison: Confusion Matrix for EfficientNet V2 B0

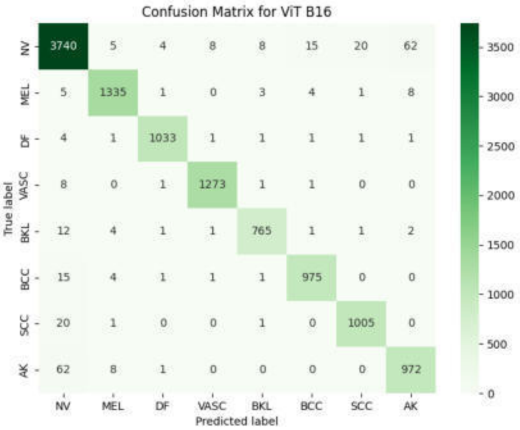


True label \ Predicted label	NV	MEL	DF	VASC	BKL	BCC	SCC	AK
NV	3847	2	3	1	2	2	2	3
MEL	2	1346	2	2	2	1	1	1
DF	1	0	1038	1	1	0	2	0
VASC	1	1	1	1278	1	0	1	1
BKL	1	2	1	1	779	1	1	1
BCC	0	1	1	0	1	994	0	0
SCC	1	1	1	2	2	0	1019	1
AK	1	1	2	1	1	1	1	1035

Fig. 5.3: Model Comparison: Confusion Matrix for VitB16

5.5. Classification Report. To offer a comprehensive analysis of performance indicators, including precision, recall, F1-score, and support, Table 5.3 is included, providing a detailed breakdown of these metrics by class. The table presents comprehensive performance metrics for each class across three models, namely EfficientNet-V2 B0, ViT B16, and the EffiViT Ensemble Model. The NV class, which exhibits the highest level of support with a total of 3862 instances, has outstanding precision and recall rates of 99.6% and 99.7% respectively when employing the Ensemble Model. These results indicate the model's remarkable capability in precisely identifying true positives and its reliability in effectively separating the NV class from other classes. Regarding Melanoma (MEL), all models exhibit elevated metrics; however, the Ensemble Model crosses the 99% barrier in precision and recall, indicating a noteworthy decrease in both false positives and false negatives for this crucial category. The performance of Dermatofibroma (DF) and Vascular lesions (VASC) is especially noteworthy, as the Ensemble Model demonstrates somewhat superior recall for DF and precision for VASC. This suggests that the Ensemble Model excels in accurately diagnosing these less common disorders.

The Ensemble Model demonstrates higher precision and memory rates for Benign keratosis-like lesions



	NV	MEL	DF	VASC	BKL	BCC	SCC	AK
NV	3740	5	4	8	8	15	20	62
MEL	5	1335	1	0	3	4	1	8
DF	4	1	1033	1	1	1	1	1
VASC	8	0	1	1273	1	1	0	0
BKL	12	4	1	1	765	1	1	2
BCC	15	4	1	1	1	975	0	0
SCC	20	1	0	0	1	0	1005	0
AK	62	8	1	0	0	0	0	972

Fig. 5.4: Model Comparison: Confusion Matrix for EffiViT

Table 5.3: Class-Wise Performance Metrics of the Models

Classes	EfficientNet-V2 B0			ViT B16			Ensemble Model: EffiViT			Support
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
NV	96.3	96.4	96.3	96.7	96.8	96.7	99.8	99.6	99.7	3862
MEL	98	98	98	98.3	98.3	98.3	99.4	99.1	99.2	1357
DF	98.7	98.7	98.7	99.1	99	99	98.9	99.5	99.2	1043
VASC	98.9	98.9	98.9	99.1	99.1	99.1	99.3	98.9	99.4	1284
BKL	97.1	96.5	96.8	98	97.2	97.6	98.7	99.6	98.8	787
BCC	96.9	96.9	96.9	97.7	97.7	97.7	99.4	99.6	99.5	1001
SCC	96.8	96.8	96.8	97.7	97.8	97.8	99.2	99.2	99.2	1032
AK	94	94	94	93	93.1	93.1	99.3	99.2	99.2	1043
Micro Avg	96.9	96.9	96.9	97.3	97.3	97.3	99.4	99.4	99.4	-

(BKL), Basal cell carcinoma (BCC), and Squamous cell carcinoma (SCC), with a notable emphasis on BCC. Specifically, the precision rate for BCC reaches 99.6%, while the recall rate reaches 99.5%. The Ensemble Model demonstrates enhanced skill in accurately discerning Actinic Keratoses (AK), a condition that often poses a substantial challenge. It achieves precision and recall rates of 99.3%, indicating its effectiveness in improving the diagnosis of this class.The row labelled 'Micro Avg' presents the mean performance measure across all classes, so providing a comprehensive performance evaluation for each model. The Ensemble Model exhibits a micro-average accuracy, precision, recall, and F1-score of 99.4%, hence illustrating its uniformity across all classes.

The high precision and recall seen across all classes demonstrate the reliability and resilience of the Ensemble Model, which are essential qualities for its potential therapeutic application. The instructive nature of the accuracy rate notwithstanding, it fails to provide a comprehensive understanding of the specific aspects of categorization. To enhance comprehension of classification performance, the confusion matrix, offers a comprehensive analysis of prediction data pertaining to each category. The Ensemble Model demonstrates excellent performance compared to other models, with an impressive accuracy rate of 99.4%. This highlights its exceptional capacity to accurately classify skin cancer in the context of diagnosis.

The comprehensive information provided by the confusion matrix analysis is crucial for comprehending the capabilities of each model in accurately categorizing particular skin lesions, hence directing clinical approaches for the screening of skin cancer. The exceptional performance of the Ensemble Model, with an accuracy rate of 99.4%, represents a notable advancement compared to current models. This achievement has the potential to redefine the benchmarks for accuracy in the categorization of skin cancer.

5.6. ROC-AUC Curve Analysis. In addition to the confusion matrix analysis, the ROC-AUC curve analysis will be a pivotal component of the evaluation. It offers a comprehensive measure of each model's performance across various threshold settings, further confirming their diagnostic reliability and clinical applicability. The ROC curves reported in this study (as shown in Figures 5.5,5.6,5.7) depict the performance evaluation of three distinct machine learning models, namely EfficientNet V2 B0, EffiVit, and ViT B16. The plotted curves illustrate the relationship between the true positive rate (TPR) and the false positive rate (FPR), allowing for an examination of the classifiers' diagnostic capabilities at different threshold levels. The EfficientNet V2 B0 model demonstrates high discriminating capabilities, as seen by its AUC values approaching 1 across all categories. This suggests a remarkable proficiency in distinguishing between positive and negative classes. The EffiVit model exhibits notable performance, as evidenced by some categories achieving an AUC of 1, suggesting the possibility of achieving complete classification accuracy. Finally, the Receiver Operating Characteristic (ROC) curve of the ViT B16 model also exhibits elevated Area Under the Curve (AUC) values, so validating the model's strong precision in tasks related to classification. The persistent positioning of these curves in close proximity to the upper left corner of the graph area signifies a notable degree of precision in forecasting, accompanied by a minimum occurrence of both false positives and false negatives. This highlights the resilience of these models in effectively carrying out their respective predictive functions. Moreover, the models exhibit high AUC values, indicating their suitability for implementation in distinct clinical contexts, such as diagnostic imaging or patient risk assessment, where achieving high levels of sensitivity and specificity is of utmost importance.

The ROC curve is an essential tool for evaluating the trade-off between sensitivity (or TPR) and specificity (1 - FPR) across different thresholds without requiring an arbitrary classification threshold. This makes the ROC curve particularly valuable in medical diagnostic tests where the cost of false negatives varies significantly with the clinical context. This perfect score on the AUC indicates that the model can discriminate perfectly between the positive and negative classes without any overlap. The high AUC values reinforce the potential of these models to act as reliable decision-support tools in medical diagnostics, potentially reducing the cognitive load on healthcare professionals and increasing diagnostic accuracy.

This finding underscores the capacity of these classifiers to serve as decision-support instruments in the field of medical diagnostics, enhancing the proficiency of healthcare professionals. In a comparative analysis, these models demonstrate comparable or superior performance to existing benchmarks in the realm of automated diagnosis, signifying a notable progression in the field of artificial intelligence within the healthcare domain.

5.7. Comparison with State of Art Models. This section provides a comparative analysis of proposed skin lesion classification models using the ISIC 2019 dataset, which is widely recognized as a crucial benchmark in the field of dermatological machine learning research. The Ensemble model, EffiViT, demonstrates superior performance compared to current benchmarks, attaining an overall accuracy rate of 99.4%. The achieved accuracy surpasses the average accuracy of 94.6% reported in the literature for the identical dataset. The individual models, namely EfficientNet-V2 B0 and ViT B16, exhibit strong performance, surpassing the reported average accuracies. The findings of this study highlight the efficacy of integrating sophisticated structures and Ensemble approaches, establishing them as promising instruments for practical implementation in the field of skin lesion diagnosis. Figure 10 visually presents the mentioned findings, emphasizing the superior precision of the models compared to those investigated in the existing literature.

Figure 5.8 illustrates the comparison of accuracy significance of proposed methodological breakthroughs in the classification of skin lesions. Significant progress has been gained in the field of diagnostic precision by utilizing advanced architectures and employing Ensemble techniques, thereby establishing a new standard. The ramifications of these breakthroughs are significant, since they have the potential to greatly enhance diagnostic outcomes and patient treatment within the field of dermatology.

6. Conclusions. In this study, we introduced EffiViT, an Ensemble Learning-Based Decision Support System for skin cancer diagnosis, emphasizing an intuitive Human-Machine Interface (HMI). The system combines the strengths of EfficientNet and Vision Transformer to achieve a classification accuracy of 99.4%, with a user interface tailored for ease of use and interpretability by medical professionals. Our findings underscore the critical role of user-centered HMI design in facilitating the clinical adoption of AI-based diagnostic tools, improving decision-making and patient outcomes in dermatology. Future efforts will focus on enhancing the HMI

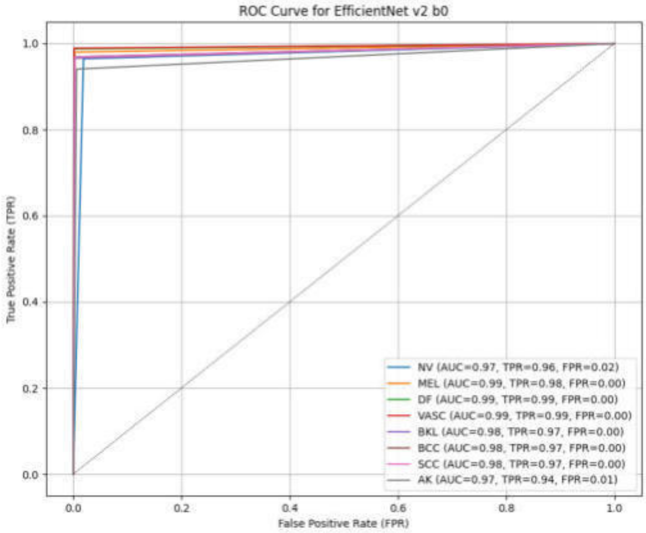


Fig. 5.5: ROC-AUC Curve for EfficientNet V2 B0

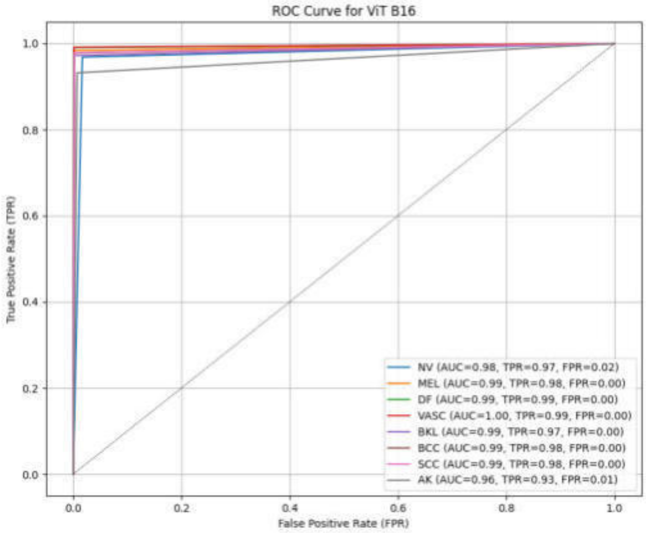


Fig. 5.6: ROC-AUC Curve for VitB16

with more interactive features and extending the system's diagnostic capabilities. By continuing to prioritize the integration of advanced technology with user-friendly interfaces, we aim to further solidify the position of AI as an invaluable asset in healthcare. The success of EffiViT illustrates the transformative potential of combining cutting-edge AI with thoughtful interface design, marking a significant step forward in human-machine collaboration in medical diagnostics.

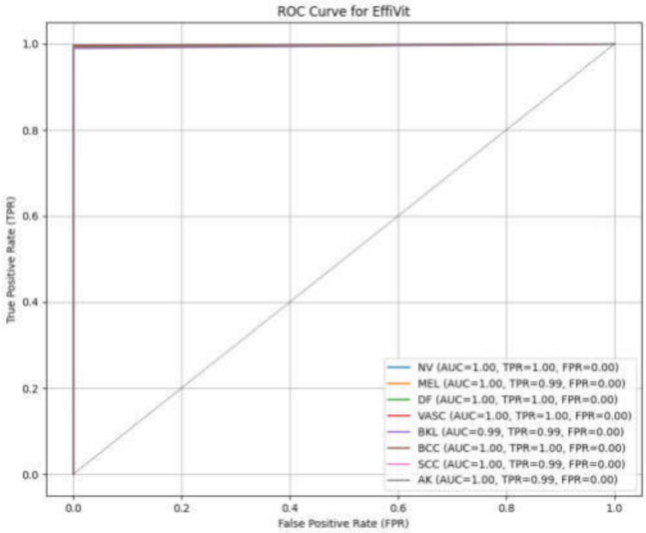


Fig. 5.7: ROC-AUC Curve for EffiViT

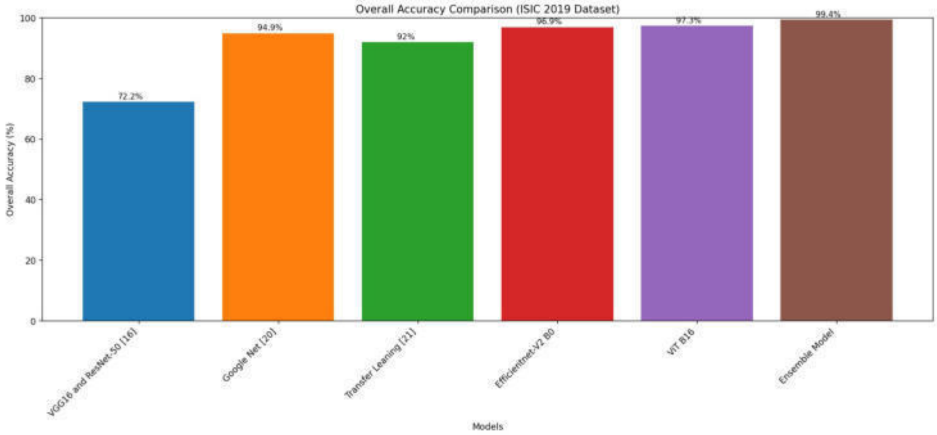


Fig. 5.8: Overall Accuracy Comparison (ISIC 2019 Dataset)

REFERENCES

[1] Andre Esteva, Brett Kuprel, and Sebastian Thrun, *Deep networks for early-stage skin disease and skin cancer classification*, Project Report, Stanford University, 2015.

[2] American Cancer Society, *Cancer Facts & Figures 2022*, American Cancer Society, Atlanta, 2022.

[3] Cancer.Net, *Melanoma: Statistics*, `https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/keystatistics.html`, Accessed 23 June 2023, 2023.

[4] Pedro MM Pereira and others, *Dermoscopic skin lesion image segmentation based on Local Binary Pattern Clustering: Comparative study*, Biomedical Signal Processing and Control, vol. 59, 101924, 2020.

[5] Sammy V. Militante, Nanette V. Dionisio, and Brandon G. Sibbaluca, *Pneumonia detection through adaptive deep learning models of convolutional neural networks*, 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), IEEE, 2020.

[6] Qing Lyu and others, *Cine cardiac MRI motion artifact reduction using a recurrent neural network*, IEEE transactions on medical imaging, vol. 40, no. 8, pp. 2170-2181, 2021.

[7] Sarfaraz Hussein and others, *Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches*, IEEE transactions on medical imaging, vol. 38, no. 8, pp. 1777-1787, 2019.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.

[9] Karen Simonyan and Andrew Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, 2014.

[10] M. Manandhar, S. Hawkes, K. Buse, E. Nosrati, and V. Magar, *Gender, health and the 2030 agenda for sustainable development*, Bulletin of the World Health Organization, vol. 96, no. 9, pp. 644, 2018.

[11] R. Erol, *Skin Cancer Malignancy Classification with Transfer Learning*, University of Central Arkansas, Conway, AR, 2018.

[12] T. Chowdhury, A. R. S. Bajwa, T. Chakraborti, J. Rittscher, and U. Pal, *Exploring the correlation between deep learned and clinical features in melanoma detection*, Annual Conference on Medical Image Understanding and Analysis, Springer, Cham, Switzerland, pp. 3-17, 2021.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, Nature, vol. 542, no. 7639, pp. 115-118, 2017.

[14] X. Li, J. Wu, E. Z. Chen, and H. Jiang, *From deep learning towards finding skin lesion biomarkers*, Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2797-2800, 2019.

[15] F. Nunnari, M. A. Kadir, and D. Sonntag, *On the overlap between grad-CAM saliency maps and explainable visual features in skin cancer images*, International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, Switzerland, pp. 241-253, 2021.

[16] M. Sadeghi, P. K. Chilana, and M. S. Atkins, *How users perceive content-based image retrieval for identifying skin images*, Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Springer, Cham, Switzerland, pp. 141-148, 2018.

[17] Y. Xie, J. Zhang, Y. Xia, and C. Shen, *A mutual bootstrapping model for automated skin lesion segmentation and classification*, IEEE Transactions on Medical Imaging, vol. 39, no. 7, pp. 2482-2493, 2020.

[18] J. Yang, F. Xie, H. Fan, Z. Jiang, and J. Liu, *Classification for dermoscopy images using convolutional neural networks based on region average pooling*, IEEE Access, vol. 6, pp. 65130-65138, 2018.

[19] H. Zunair and A. B. Hamza, *Melanoma detection using adversarial training and deep transfer learning*, Physics in Medicine & Biology, vol. 65, no. 13, Article 135005, 2020, doi: 10.1088/1361-6560/ab86d3.

[20] S. H. Kassani and P. H. Kassani, *A comparative study of deep learning architectures on melanoma detection*, Tissue & Cell, vol. 58, pp. 76-83, 2019.

[21] M. A. Kassem, K. M. Hosny, and M. M. Fouad, *Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning*, IEEE Access, vol. 8, pp. 114822-114832, 2020.

[22] J. A. A. Salido and C. Ruiz, *Using deep learning for melanoma detection in dermoscopy images*, International Journal of Machine Learning and Computing, vol. 8, no. 1, pp. 61-68, 2018.

[23] A. H. Shahin, A. Kamal, and M. A. Elattar, *Deep ensemble learning for skin lesion classification from dermoscopic images*, Proceedings of the 9th Cairo International Biomedical Engineering Conference (CIBEC), pp. 150-153, 2018.

[24] F. Sherif, W. A. Mohamed, and A. Mohra, *Skin lesion analysis toward melanoma detection using deep learning techniques*, International Journal of Electronics and Telecommunications, vol. 65, no. 4, pp. 597-602

[25] H. M. Ünver and E. Ayan, *Skin lesion segmentation in dermoscopic images with combination of Yolo and GrabCut algorithm*, Diagnostics, vol. 9, no. 3, Article 72, 2019.

[26] M. Tan and Q. Le, *EfficientNetV2: Smaller Models and Faster Training*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[27] Y. Wang, X. Ma, Z. Chen, and others, *Symmetric cross entropy for robust learning with noisy labels*, Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, pp. 322-330, 2019.

[28] A. Paszke, S. Gross, F. Massa, and others, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, Advances in Neural Information Processing Systems, vol. 32, pp. 8024-8035, 2019.

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Mar 16, 2024
Accepted: May 27, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 84–95, DOI 10.12694/scpe.v26i1.3574

SECURE DIGITAL DATA VISUAL SHARING SCHEMES IN MULTI-OWNER PUBLIC CLOUD ENVIRONMENT APPLICATIONS

CHEKKA RATNA BABU*AND B. RAVEENDRA BABU †

Abstract. The use of cloud computing for storing and processing cyber data via the internet has grown widespread in the field of cyber security. Ensuring the secure sharing of cyber data, including texts, images, audio, and video, over the cloud is paramount. However, cloud computing encounters significant challenges concerning cyber data security, authentication, and privacy. The prevailing approaches to data security encounter challenges such as the generation of intricate keys, intensive computation for large keys, and susceptibility to attacks by intruders. Scalability presents its own set of obstacles in maintaining cyber data privacy and ensuring secure communication. A significant privacy concern arises from the frequent changes in membership and data sharing among multiple owners. This paper introduces a novel approach to secure data sharing within data centres by proposing a scheme that employs Private Key Dynamic Visual Cryptography (PK-DVC), Multiple Key Encryption Visual One-Time Pad (MK-VOTP), and visual steganography for encoding and decoding. MK-VOTP is used for data encryption, facilitating data owners' encryption via utilising their identity with supplementary security features. Subsequently, the encrypted data is kept in the cloud, guaranteeing heightened security protocols. Visual steganography is employed for further authentication purposes. Decrypting the original data requires users who fulfil the encrypted properties. This increases security and reduces critical size, allowing several authenticated owners to share data without conflicts. Combining all shares recreates the original picture.

The approach described in the paper sounds innovative and addresses several key challenges in ensuring secure data sharing within data centers, particularly in the context of cloud computing.The proposed scheme main components are *Private Key Dynamic Visual Cryptography (PK-DVC)*,*Multiple Key Encryption Visual One-Time Pad (MK-VOTP)*, and *Visual Steganography*.These three components plays important role in further authenticating the encrypted data or providing additional security measures beyond encryption.By combining these techniques, the proposed scheme aims to provide robust security for sharing cyber data in cloud environments. The use of dynamic keys, multiple encryption layers, and steganography can enhance data security, making it challenging for intruders to access or decipher sensitive information.

Key words: Multi-Owner, Encryption, Visual Cryptography, Decryption, Cloud Computing, Visual Steganography

1. Introduction.

Since 2007, the IT industry and academics have focused on cloud computing. The idea involves assigning services in a "cloud," a network of devices and resources linked via the Internet, to seamlessly offer more flexible services to consumers. Data storage, an essential component of cloud computing, presents new issues in developing safe and reliable storage and access methods among cloud service providers [5]. Given cloud computing's inherent openness, cloud data storage must be secured [6].In Figure.1 Multiowner − Cloud Architecture Access control is crucial in cloud-based systems due to data accessibility.

However, ensuring the security and privacy of data stored in the cloud presents significant challenges. While data encryption offers protection, concerns arise when cloud servers possess decryption keys and manage user accessibility rights, particularly for sensitive records like administrative documents (e.g., bills and ID cards)[5]. Additionally, as users increasingly outsource storage and computing needs to remote servers, often untrusted, privacy issues escalate, including the sensitivity of keywords transmitted in queries and the retrieved data, requiring concealment. Consequently, security and privacy emerge as paramount concerns. Traditional network security and privacy mechanisms need to be revised or address data storage and outsourcing complexities. Maintaining the integrity and confidentiality of stored digital data identities becomes a significant challenge external storage solutions pose [6]. Dynamic visual cryptography and steganography techniques mitigate these risks and preserve data privacy and authentication in the cloud.

Dynamic Visual Cryptography: The dynamic visual cryptography scheme offers a robust approach to encrypting confidential documents or images by partitioning them into distinct segments. One noteworthy

[*] Acharya Nagarjuna University, Guntur, India (chekka.ratnababu@gmail.com)
[†] Accreditations & Recruitment, Vishnu Group of Institutions, Hyderabad & Bhimavaram, India (rbhogapathi@yahoo.com)

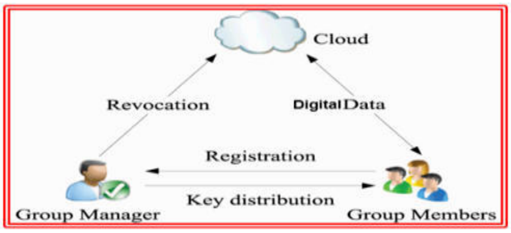


Fig. 1.1: Multiowner - Architecture

characteristic of visual cryptography [7][8][9] is the ability to visually decipher the secret picture by superimposing shares, hence obviating the need for intricate computational processes. Exploiting this property, a third party can easily reconstruct the secret image if the shares are transmitted sequentially over the network. This approach involves encrypting visually generated image shares using Private Key Encryption, with the primary objective of achieving a high level of security. Unlike previous systems where only a single user could access the data, this system enables multiple users to access the same data. Moreover, it utilizes more minor keys than the larger keys previously employed. Addressing the paramount concern of security in data sharing over the cloud, this system aims to mitigate associated risks effectively.

2. Relade Work. Data undergoes encryption before outsourcing, resulting in the service provider receiving encrypted data. Consequently, this data is perceived as useless or lacking in meaning. However, the responsibility for managing access control policies, encrypting and decrypting data, and managing cryptographic keys falls on the client [10]. Adding to the user's burden and sharing it increases hazards. Data shared among several users requires more encryption flexibility to accommodate group members, enable key management, and enforce access control regulations to protect data confidentiality. Outsourced data owners have extra constraints when sharing data with many consumers.

Public key encryption encounters difficulties when implemented in cloud environments, with numerous users needing file access. Sana et al. introduced a lightweight encryption algorithm in their study cited as [11], melding symmetric encryption for file encryption and asymmetric encryption for crucial distribution to tackle these challenges. Nonetheless, this approach has drawbacks, including complexities in key management and the necessity for precise file access control, as noted in [12]. Furthermore, the solution's flexibility and scalability could be improved. Encryption and decryption [29] methods must be implemented whenever a user leaves a group to prevent unauthorised access to data.

Shamir established the concept of identity-based encryption [13]. This encryption method involves data owners encrypting their data by providing the identity of the authorised entity responsible for decrypting it. The decryption entity's identity must correspond to the one designated by the owner, therefore obviating the need for key exchange.

Attribute-based encryption (ABE) is a cryptographic technique that establishes user identification [32] using a characteristic collection. These qualities are then used to build a secret key and establish the access structure for access control. This approach integrates encryption with access control, allowing for data confidentiality and sharing among groups of users. In [14], fuzzy identity-based encryption (FIBE), a variant of ABE, was proposed several years after IBE. In FIBE, a group of attributes collectively identifies an individual's identity. The data owner encrypts the data, and only individuals possessing attributes that overlap with those specified in the ciphertext can decrypt it.

Ostrovsky's technique [15] introduces a method that enables the implementation of non-monotonic access structures, which may accommodate both positive and negative qualities. Nevertheless, this technique leads to a rise in the size of the ciphertext and the key and the accompanying time expenses for encryption and decryption. On the other hand, Key-Policy Attribute-Based Encryption (KP-ABE) exhibits a linear growth in the ciphertext size as the number of related characteristics increases.

An algorithm has been devised to guarantee a consistent ciphertext size irrespective of the number of characteristics while accommodating non-monotonic access patterns. Nevertheless, the dimensions of the critical exhibit a quadratic growth pattern as the number of factors increases. Cypher Text Policy Attribute-Based Encryption (CP-ABE) was developed to address this problem [16]. Nevertheless, it is essential to acknowledge that CP-ABE often entails more expenses than KP-ABE.

Cybersecurity between Users W. Chen and Wang's Mobile Media Cloud [17] secured multimedia data across users' clouds. The system used DCT RS code watermarking, picture concealing, and secret sharing. Its advantages were strong security, user-friendliness, media quality, and low overhead. One downside was that picture data was limited to one carrier. This constraint is overcome by dividing the picture into shares and transmitting the data over carriers.

S, Chauhan, and Vats suggested Threshold Cryptography-Based Data Security in Cloud Computing [18] to protect communication between Data Owners (DO), Managed Service Providers (MSP), and cloud customers and cyber data access. The system used Diffie-Hellman, threshold cryptography, and MD5. Its benefits were reduced keys and improved data security, confidentiality, integrity, and performance. Diffie-Hellman algorithm computing power was a system downside. To address this issue, the system replaced the Diffie-Hellman algorithm with MK-VOTP, a lightweight encryption method, to achieve efficient encryption and decryption without requiring extensive computational power.

The goal of the Digital Image Sharing by Diverse Image Media [19][30] method proposed by Lee and Chiu was to partition photographs into several shares safely. The system utilized techniques such as the NSS algorithm and encoding algorithms. Its benefits encompassed user-friendliness, high security, quality images, and low-risk transmission. However, a drawback of the system was generating noise-like shares during the process. Mukherjee and Ghoshal presented ”Steganography-based Visual Cryptography” [20] to bolster image security by integrating Steganography with Visual Cryptography (VC), a method where data is concealed within another image. The system employed techniques such as cryptography and steganography. Its advantages included heightened security, authentication, integrity, and reduced computational overhead. However, a potential drawback could be a decline in image quality. In this study, efforts were made to address weaknesses such as poor image quality and noisy generated shares while enhancing privacy, authentication, and robustness.

3. Proposed Work. Securing information to restrict access solely to authorized individuals has been a longstanding practice. However, this pursuit encounters challenges in both physical and digital domains. Even well-protected information remains susceptible to theft or unintended misuse in the tangible world. Similarly, the cyber realm faces comparable obstacles, often resorting to container-based encryption for safeguarding. The analogy of lock-and-key security persists in digital landscapes, where data integrity hinges on robust encryption techniques. Within cloud environments[21], digital identity is the cornerstone of adaptable data security. Digital identity encompasses diverse attributes defining an individual, thereby presenting risks of identity theft. Consequently, upholding the confidentiality of digital identities is crucial, not only for security but also for privacy.

For instance, email represents a typical scenario of multi-owner privacy data[2]. When two individuals exchange an email, the content becomes a shared privacy concern and which is shown in Figure 3.1. The contents should not be disclosed without either party's consent [3][4]. Nevertheless, given that the email is owned by both the correspondent and the recipient[1], they each possess the authority to forward it autonomously. Forwarding entails the inherent danger of revealing the confidential data of the other proprietor.

The Online Patient Health Record System (OPHRS) looks promising for online patient health information exchange. Patients may manage, regulate, and share their health data with other users and doctors. When OPHRS is semi-trusted to a third-party server, issues about unauthorised access, privacy breaches, and security vulnerabilities develop, creating substantial problems in a multi-owner cloud environment. A secure cloud-based OPHRS architecture addresses these issues. MK-VOTP and visual steganography enable safe OPHRS sharing across several users in this system.

The proposed system is a collaborative platform to facilitate secure image sharing among multiple users on cloud infrastructure. Given the inherent limitations of cloud services in ensuring security, confidentiality, and integrity, this system employs Visual Cryptography[22][23] to distribute image components among group members. Each segment of the image undergoes encryption and decryption procedures using dynamic visual cryptography. Visual steganography[24][25] is integrated as an additional protective layer to fortify security measures and validate authentication. Traditionally, images were transmitted as single carriers, leaving them

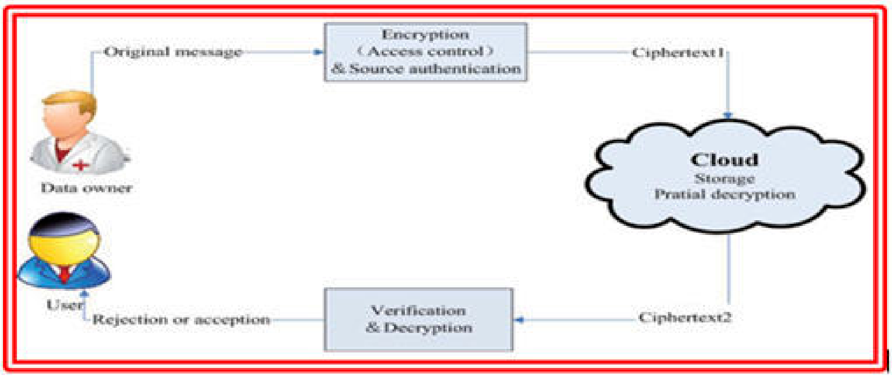


Fig. 3.1: Cloud Environment Encryption/Decryption

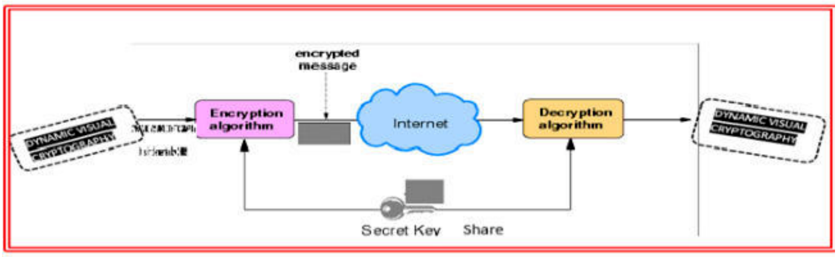


Fig. 3.2: Private Key Dynamic Visual Encoding/Decoding

vulnerable to loss. A single image is fragmented into multiple shares and carriers to mitigate this risk. Cloud service users must understand computing environment obligations and security and privacy concerns. Thus, the suggested system provides cloud computing solutions for organisational security and privacy. The ultimate aim of the proposed system is to achieve real-time image transfer to authenticated users. By consolidating the image shares, the complete image is reconstructed. Any unauthorized alterations to the image trigger access restrictions for users. The system effectively reduces the likelihood of data tampering or unauthorized access through multiple encryption and decryption operations.

3.1. Private key Dynamic Visual Cryptography Encryption. Dynamic Visual Cryptography [26-28] is an exceptional encryption technique designed to hide information within images, enabling decryption by human vision with the correct key image. This method employs two transparent images: one containing random pixels (the key) and the other holding the secret information. Retrieving the hidden data from either image alone is unfeasible; both are essential to reveal the concealed information. When the random image comprises genuinely random pixels, it functions similarly to a One-time Pad system, delivering encryption highly resistant to decryption.All algorithmic steps are revealed in Figure 3.2.

3.2. Multiple Key Encryption Visual One–Time Pad. MK-VOTP is suggested for cloud owner/user authentication. Figure 3.3 solution allows several owners to authenticate to the cloud server using multi-level security. Several owners may obtain shares from the group management and utilise the token several times.

Since the cloud cannot guarantee security, secrecy, or integrity, this multi-owner solution lets anybody safely transmit photos to numerous owners on the cloud. Each group member distributes a portion of the picture using Dynamic Visual Cryptography [33][34]. Dynamic visual cryptography encrypts and decrypts each picture sharing. Steganography will be used to improve security and check for sharing manipulation. Organisations and people using cloud services must comprehend computing environment obligations and security and privacy implications. Therefore, the suggested method ensures that cloud computing solutions meet organisational

Algorithm 1 Private Key Dynamic Visual Encoding/Decoding

a.Generate Private Key(imgKey):

```
Input:The given input consists of a Gray-Scale Image (GI) with t1 x t2 pixels dimensions.
 The basis matrices C0 and C1 have sizes of n x m.
Output: Securely generates a new Key Image
Step 1: First, create a fully transparent white pixel grey scale (GI) image with size t1 × t2 pixels
Step 2:  an image to transform into a key by turning its black pixels into 3/4 pixel blocks
Step 3:  an image to transform into a key by turning its white pixels into 2/4 pixel blocks
            (which itself are randomly determined).
Step 4: Non-white pixels are treated as if they were black.
Step 5: Fully white image for truly random key
```

b.Encoding(imgKey, imgSrc) Pseudo Code:

```
Step 1:BufferedImageencryptedImage=new BufferedImage(key image.getWidth(),key image.getHeight(),
           BufferedImage.TYPE_INT_ARGB);
Step 2: for (int y = 0; y < encryptedImage.getHeight(); y += 2) {
Step 3: for (int x = 0; x < encryptedImage.getWidth(); x += 2) {
Step 4:   if (sourceImageRes.getRGB(x, y) == Color.BLACK.getRGB()) {
Step 5:   if (keyImage.getRGB(x, y)>>>24 == 0)
          encrGraphics.fillRect(x, y, 1, 1);
          if (keyImage.getRGB(x + 1, y)>>>24 == 0)
          encrGraphics.fillRect(x + 1, y, 1, 1);
      if (keyImage.getRGB(x, y + 1)>>>24 == 0)
          encrGraphics.fillRect(x, y + 1, 1, 1);
     if (keyImage.getRGB(x + 1, y + 1)>>24 == 0)
          encrGraphics.fillRect(x + 1, y + 1, 1, 1);
      } else {
      if (keyImage.getRGB(x, y) == Color.BLACK.getRGB())
          encrGraphics.fillRect(x, y, 1, 1);
          if (keyImage.getRGB(x + 1, y) == Color.BLACK.getRGB())
          encrGraphics.fillRect(x + 1, y, 1, 1);
   if (keyImage.getRGB(x, y + 1) == Color.BLACK.getRGB())
          encrGraphics.fillRect(x, y + 1, 1, 1);
   if (keyImage.getRGB(x + 1, y + 1)== Color.BLACK.getRGB())
          encrGraphics.fillRect(x + 1, y + 1, 1,  1);}}}
Step 6: return encryptedImage;
```

c. Decoding(keyImage, overlayImage) Pseudo Code:

```
Step 1: BufferedImage cleanImage = new
        BufferedImage(overlayImage.getWidth() / 2,  overlayImage.getHeight() / 2 , BufferedImage.TYPE_INT_ARGB);
Step 2: for (int yOverlay = 0, yClean = 0; yOverlay <overlayImage.getHeight(); yOverlay += 2,  ++yClean)
Step 3: for (int xOverlay = 0, xClean = 0; xOverlay <overlayImage.getWidth(); xOverlay += 2,  ++xClean) {
             int rgbFirstPixel = overlayImage.getRGB(xOverlay, yOverlay);
Step 4: if (rgbFirstPixel >>>24 != 0 &&
        overlayImage.getRGB(xOverlay + 1, yOverlay) >>>24 != 0  &&
         overlayImage.getRGB(xOverlay, yOverlay + 1) >>>24 != 0   &&
         overlayImage.getRGB(xOverlay + 1, yOverlay + 1) >>>24 != 0) {
cleanGraphics.setColor(new Color(rgbFirstPixel, true));
cleanGraphics.fillRect(xClean, yClean, 1, 1);
}}}
Step 5: return cleanImage;
```

security and privacy needs.

The ultimate result of the suggested system is the successful real-time delivery of the secret picture to the authorised users. The original image will be created by merging the shared portions of the image. If any manipulation is performed on the picture, the user's ability to access the image will be limited. The picture undergoes various encryption and decryption processes, making tampering with or hijacking the data challenging. Therefore, if an unauthorised user attempts to get access to the confidential picture, they will be unable to do so if any malicious user tampering occurs with the secret image since it may be discovered via steganography. If any user's private key is absent, the original picture will remain inaccessible to all users.

3.3. Secure Secret Share Visual Steganography.

A network of interconnected bank branches provides cloud banking services [37]. Cloud bank customers can access their assets and carry out basic transaction

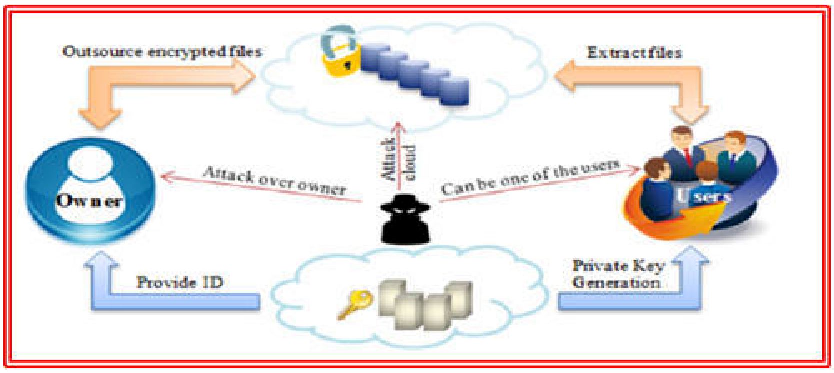


Fig. 3.3: Multiple Key Encryption Visual One–Time Pad

Algorithm 2 Algorithm Multiple Key Encryption Visual One–Time Pad Pseudo Code:

```
EncDoubledKey(Image newImage1, Image newImage2, int height, int width, int n){
Step 1: m_enc2 = new Enc2_2(height, width, n);
Step 2: m_loadingImage1 = newImage1;
  m_loadingImage2 = newImage2;
Step 3: this.initEncrypt();
       }
               initEncrypt(){
        Step 3.1: m_enc2.doPermutation();
        Step 3.2: m_enc2.setImage(m_loadingImage2);
        Step 3.3: m_enc2.encrypt();
                       //this.doPermutation();
        Step 3.4: m_Cblack = m_enc2.m_Cblack;
        Step 3.5: m_Cwhite = m_enc2.m_Cwhite;
        Step 3.6: this.setImage(m_loadingImage1);
        Step 3.7: this.encrypt();
        return true;  }
```

activities from any member cloud network bank branch office. One of the primary challenges encountered in cloud banking services is the integrity of the members. The authentication of information in the cloud banking industry is a significant challenge due to the occurrence of SQL Injection hacking attacks on bank databases. To address the challenges associated with authentication, this study proposes a method that utilises image processing techniques to secure secret sharing using visual steganography. The present study introduces a novel approach to embedding a customer's PIN using secure secret share visual steganography. While many existing steganography methods rely on three or four adjacent pixels over a single pixel, the secure secret share visual steganography technique [35][36] offers the advantage of utilising up to eight adjacent neighbours. This allows for a gradual increase in value, which can then be divided into multiple segments. The quantity of shares to be generated is contingent upon the specific plan the financial institution selects. After creating two shares, one share is recorded in the bank database, and the client retains the second share. The consumer needs to ensure the presentation of their share throughout all transactions. The original picture is obtained by stacking this share with the first share. The decoding process extracts the concealed pin number upon acknowledging or rejecting the output, verifying the customer's identity.

Multi – Owner Public Cloud Environment is a conceptual model that outlines the cloud application, behavior, and aspects, as depicted showin Figure 3.4.

Figure 3.5.(a) shows the logic flow to generate shares by using MK-VOTP algorithm and visual Stenography algorithm. If both the keys (private key 1 and private key 2) are same, cloud member authenticity may be

Algorithm 3 Visualstegnography Algorithm

```
Step 1: BufferedImage keyFromInitialImage = Crypting.generateKey(imgFirst);
Step 2: BufferedImage secondImageCanvas =
        new BufferedImage(width*2, height * 2, BufferedImage.TYPE_INT_ARGB);
Step 3: for (int y = 0; y < height; ++y) {
        for (int x = 0; x < width; ++x) {
        int newX = x * 2;
        int newY = y * 2;
        int pixelRGB = imgToHide.getRGB(x, y);
        boolean targetShouldBeBlack = (pixelRGB >>> 24 != 0); // Check transparency
        pixelRGB = secondImageCanvas.getRGB(x, y);
        boolean secondImagePixelIsBlack = (pixelRGB >>> 24 != 0);
Step 4: int blackPixelsToSet = secondImagePixelIsBlack ? 3 : 2;
Step 5: boolean skipFirst = !targetShouldBeBlack;
        for (int minix = 0; minix < 2; ++minix) {
        for (int miniy = 0; miniy < 2; ++miniy) {
        int firstPixelRGB = keyFromInitialImage.getRGB(newX + minix, newY + miniy);
        boolean isFirstPixelWhite = (firstPixelRGB != Color.BLACK.getRGB());
        if (isFirstPixelWhite) {
          if (skipFirst) {
            skip first = false;
            newPixels[minix][miniy] = Boolean.FALSE;
            continue;
          }
          newPixels[minix][miniy] = Boolean.TRUE;
          --blackPixelsToSet; } } } } }
```

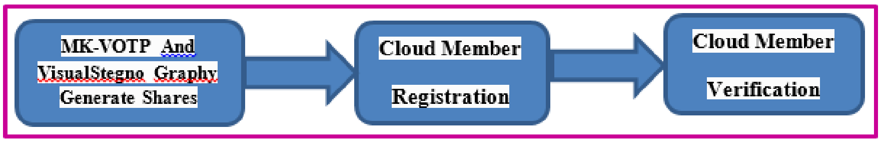


Fig. 3.4: Proposed Method Cloud Multi Owner Security Flow Diagram

granted and if both are not same, admin can decide that the share produced by cloud member is fake and can be rejected. In Cloud Member Registration/Verification to register [Figure 3.5.(b)] the cloud member and Verify (Figure. 3.5.(c)) the correct cloud member or not.

4. Experiments and Results. The outcome of the system entails the retrieval and creation of the first picture on the customer's end via the accumulation of the shares held by the participating consumers. The agent sends the image, which is then subjected to Visual Cryptography. This involves splitting the image into n shares. The sensitive data is then encrypted using private critical dynamic visual cryptography and a one-time multiple-key encryption visual pad. The encoded and decoded shares are then authenticated. Finally, Visualstegnography is used to authenticate the customers .The Simulation results shown in from Figure. 4.1a to Figure 4.2c step by step in sequence order.

The efficiency of Visual cryptosystem depends on the quality of the reconstructed image. The important parameters of proposed VC scheme are pixel expansion(m), which refers to the number of pixels in a share used to encrypt a pixel of the secret image. This implies loss of resolution in the reconstructed image and contrast (α), which is the relative difference between black and white pixels in the reconstructed image. This implies the quality of the reconstructed image. Generally, smaller the value of m will reduce the loss in resolution and greater the value α of will increase the quality of the reconstructed image. As mentioned above if 'm' is decreased, the quality of the reconstructed image will be increased but security will be a problem The secret

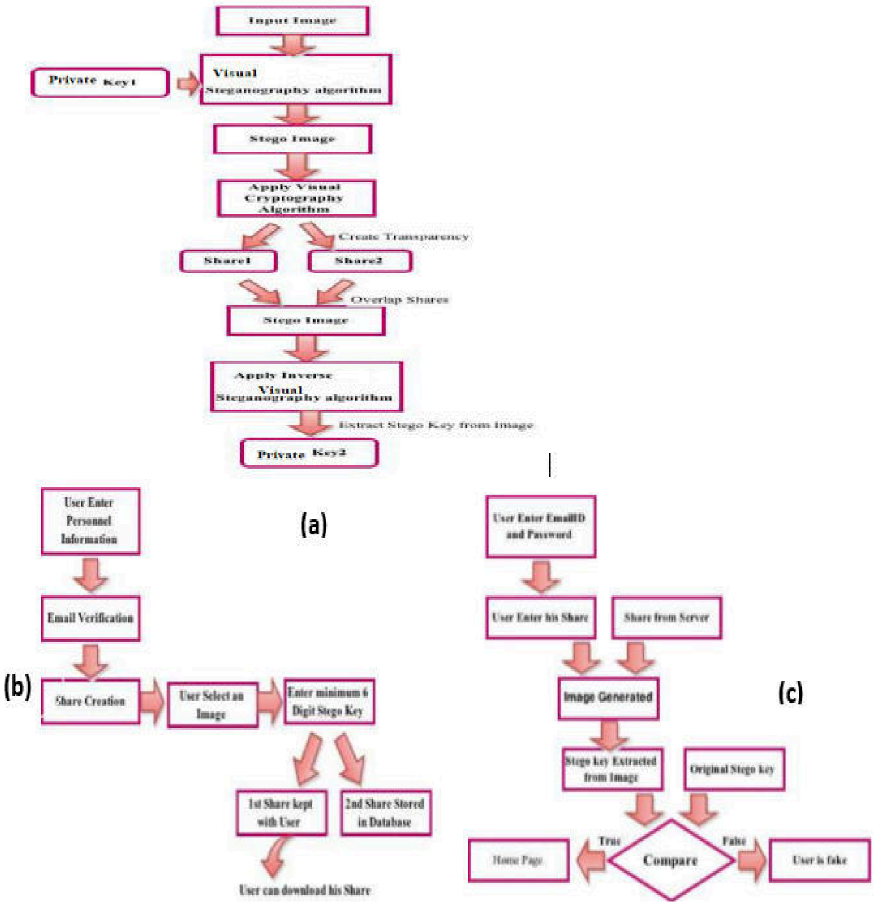
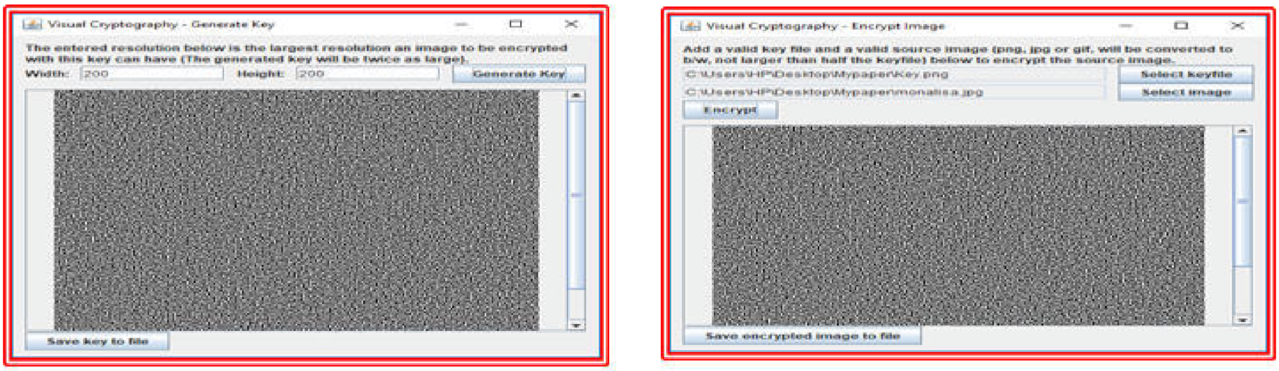

Input Image
Private Key1
Visual Steganography algorithm
Stego Image
Apply Visual Cryptography Algorithm
Create Transparency
Share1
Share2
Overlap Shares
Stego Image
Apply Inverse Visual Steganography algorithm
Extract Stego Key from image
Private Key2
(a)
User Enter Personnel Information
Email Verification
(b)
Share Creation
User Select an Image
Enter minimum 6 Digit Stego Key
1st Share kept with User
2nd Share Stored in Database
User can download his Share
User Enter EmailID and Password
User Enter his Share
Share from Server
Image Generated
(c)
Stego key Extracted from Image
Original Stego key
True
Compare
False
Home Page
User is fake

Fig. 3.5: Proposed Method Cloud Multi Owner Security Flow Diagram


Visual Cryptography - Generate Key

The entered resolution below is the largest resolution an image to be encrypted with this key can have (The generated key will be twice as large).

Width: 200 Height: 200

(a) Generate Key

Visual Cryptography - Encrypt Image

Add a valid key file and a valid source image (png, jpg or gif, will be converted to b/w, not larger than half the keyfile) below to encrypt the source image.

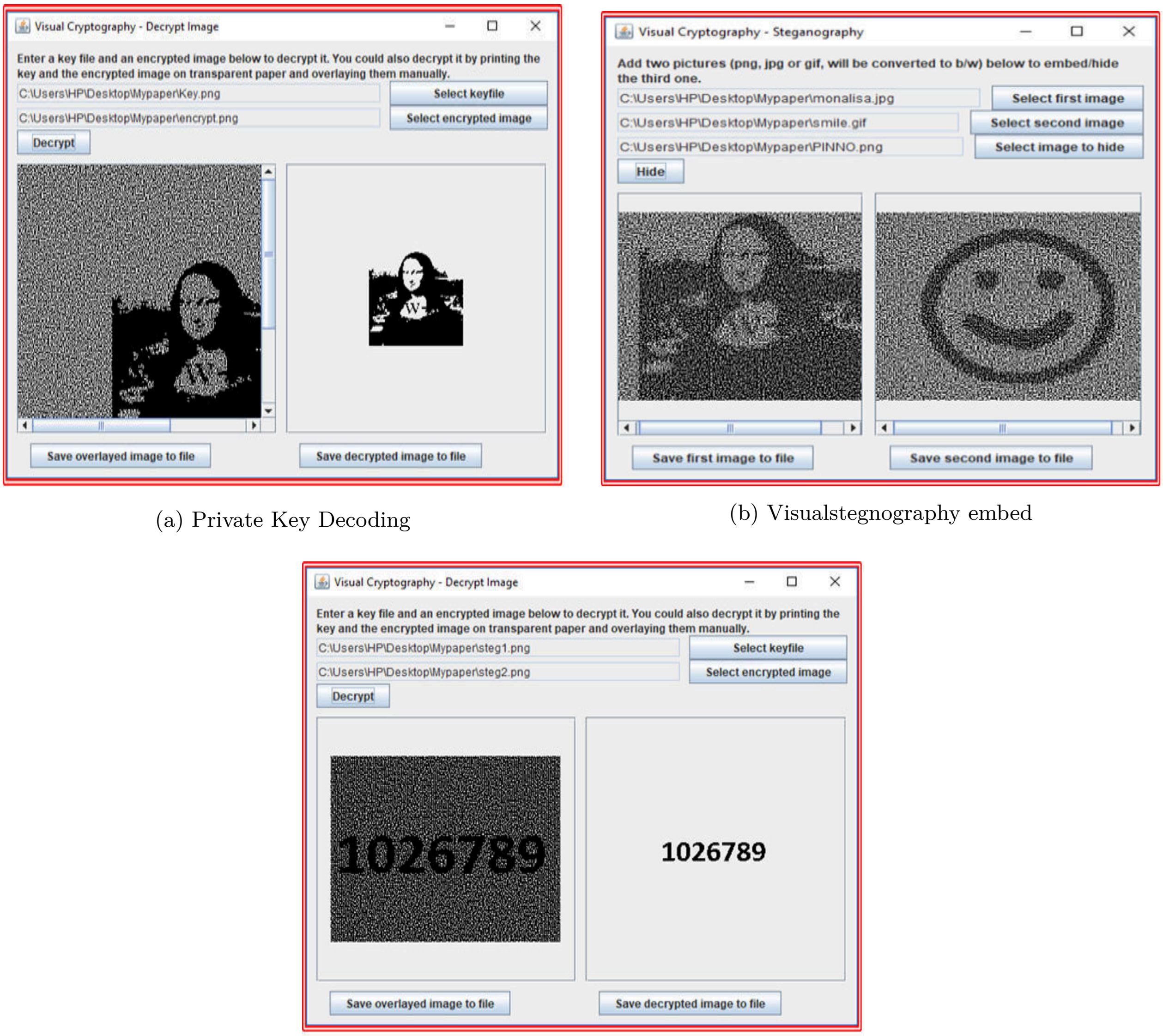
C:\Users\HP\Desktop\Mypaper\Key.png

C:\Users\HP\Desktop\Mypaper\monalisa.jpg

(b) Private Key Encoding

Fig. 4.1: Generate Key and Private Key Encoding

share or reconstructed image generated from the original image (Figure 4.3a) is shown in Figure 4.2b, the



Visual Cryptography - Decrypt Image

Enter a key file and an encrypted image below to decrypt it. You could also decrypt it by printing the key and the encrypted image on transparent paper and overlaying them manually.

C:\Users\HP\Desktop\Mypaper\Key.png

C:\Users\HP\Desktop\Mypaper\encrypt.png

Select keyfile
Select encrypted image
Decrypt
Save overlayed image to file
Save decrypted image to file

(a) Private Key Decoding

Visual Cryptography - Steganography

Add two pictures (png, jpg or gif, will be converted to b/w) below to embed/hide the third one.

C:\Users\HP\Desktop\Mypaper\monalisa.jpg

C:\Users\HP\Desktop\Mypaper\smile.gif

C:\Users\HP\Desktop\Mypaper\PINNO.png

(b) Visualstegnography embed

Enter a key file and an encrypted image below to decrypt it. You could also decrypt it by printing the key and the encrypted image on transparent paper and overlaying them manually.

C:\Users\HP\Desktop\Mypaper\steg1.png

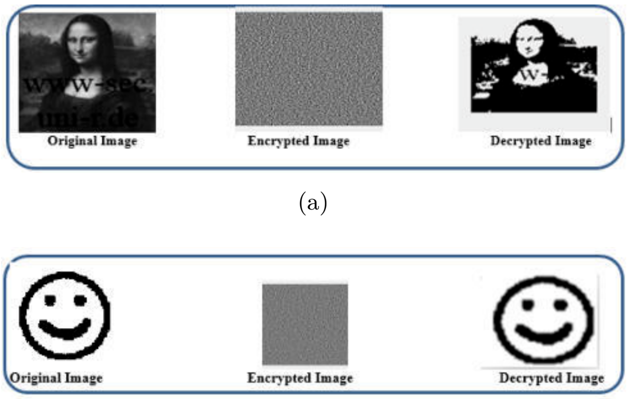
C:\Users\HP\Desktop\Mypaper\steg2.png

(c) Visualstegnography (overlayed and Descrypted Image)

Fig. 4.2

corresponding ownership share is shown in Figure 4.2c, and the stacked result of Figure 4.2a and Figure 4.2b is illustrated in Figure 4.2c. In the Secrete Share the ratio of block pixels to white pixels is 50.21 to 49.79,Which reflects the central limits statement. In addition, two common similarity measurements are introduced to evaluate the proposed cloud membership protection scheme.One is the peak signal-to-noise ratio (PSNR) used to evaluate the similarity of two grey-level images

$$PSNR = 10 \times \frac{log(255)^2}{MSE}$$



(a)

(b)

Fig. 4.3: Visualstegnography (PSNR/SSIM)

Table 4.1: Quality measures of the images

(a) For Figure 4.3a

Image	**SSIM index**	**PSNR**
Original Image	1	18.49dB
Encrypted Image	0.0004	31.79 dB
Decrypted Image	0.90	20.33 dB

(b) For Figure 4.3b

Image	**SSIM index**	**PSNR**
Original Image	1	19.54 dB
Encrypted Image	0.0003	33.38 dB
Decrypted Image	0.89	21.22 dB

where

$$MSE = \frac{1}{M1 \times M2} \times \sum_{i=1}^{M1} \sum_{j=1}^{M2} (hij - (hij^1)^2)$$

hij denotes a pixel color of the original image,and h1ij denotes a pixel color of the attacked image,$M1 \times M2$ is the size of the image. Another parameter is the Structural Similarity (SSIM) index for measuring the quality between two images. The SSIM index can be viewed as a quality measure of one of the images being compared provided the other image is regarded as of perfect quality. The quality measures are calculated between the original image and the encrypted/decrypted image. Table 4.1a and Table 4.1b shows the quality measures of the images in Figure 4.3a and in Figure 4.3b.

5. Conclusion. This paper explores a growing trend in implementing multiple security techniques to manage security risks on the cloud effectively, thus unlocking the full potential of cloud computing. The proposed approach significantly contributes to cloud security and privacy, focusing on authentication and protecting image copyright and multimedia data confidentiality. Visual cryptography is employed for encrypting, concealing, and sharing information in image form, ensuring that the information is only visible to the human eye upon decryption with the correct key. Visual steganography embeds secret information and various features into original images, facilitating the identification of ownership of modified images and addressing issues related to tampering and verification. Sensitive data is secured using dynamic visual cryptography with private keys and Multiple Key Encryption Visual One-Time Pad methods. Users may encrypt their data with their identity and other security features and save it in the cloud. Additionally, visual steganography improves authentication .

Moreover, the scheme seems to address scalability concerns by allowing multiple authenticated owners to share data without conflicts, thus ensuring secure communication even in dynamic environments with frequent changes in membership and data sharing.

Overall, this approach appears promising in mitigating the challenges associated with cyber data security, authentication, and privacy in cloud computing environments. However, its effectiveness would need to be evaluated through rigorous testing and analysis to ensure its practical viability and resilience against potential attacks.

REFERENCES

[1] Yi Ren, Fangquan Cheng, Zhiyong Peng, Xiaoting Huang, Wei Song. ”A privacy policy conflict detection method for multi-owner privacy data protection”, *Electronic Commerce Research*, 2010.

[2] F. Koufogiannis and G. J. Pappas, ”Multi-owner multi-user privacy,” *IEEE 55th Conference on Decision and Control (CDC)*, Las Vegas, NV, USA, pp. 1787-1793, 2016.

[3] Saikiran Ellambotla, Dr. Anubarti, Dr. Md.Ateeq Ur Rahman. “Cloud Computing Data Group Distribution and Restricted Distribution with Multi Owner”, *International Journal of Engineering Science Invention (IJESI)*, Volume 8, Issue 11, PP 37-45, 2019.

[4] X. Liu, Y. Zhang, B. Wang and J. Yan, ”Mona: Secure Multi-Owner Data Sharing for Dynamic Groups in the Cloud,” in *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1182-1191, June 2013.

[5] Abo-alien, A., Badr, N.L., Tolba, M.F. ”Data Storage Security Service in Cloud Computing: Challenges and Solutions”, *Multimedia Forensics and Security. Intelligent Systems Reference Library*, vol 115. Springer, Cham. 2017.

[6] Paul, V., Mathew, R. (2020). “Data Storage Security Issues in Cloud Computing”. In: Pandian, A., Palanisamy, R., Ntalianis, K. (eds) *Proceeding of the International Conference on Computer Networks, Big Data and* IoT (ICCBI - 2019). ICCBI 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 49. Springer, Cham.

[7] S. -J. Lin and W. -H. Chung, ”A Probabilistic Model of (t,n) Visual Cryptography Scheme With Dynamic Group,” in IEEE Transactions on Information Forensics and Security, vol. 7, no. 1, pp. 197-207, Feb. 2012

[8] Chitra, K., Prasanna Venkatesan, V. (2020). “A Dynamic Security Model for Visual Cryptography and Digital Watermarking”. In: Pandian, A.P., Senjyu, T., Islam, S.M.S., Wang, H. (eds) Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018). ICCBI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 31. Springer.

[9] Rajkumar Kushwaha, Ankit Gajakosh, Prasanna More, Manjusha Shelke, Nilima Patil. “Visual Cryptography”, International Journal of Creative Research Thoughts Volume 6, Issue 1, pp 584-587, 2018.

[10] Zhang, R., Wang, J., Song, Z. et al. “An enhanced searchable encryption scheme for secure data outsourcing”. Sci. China Inf. Sci. 63, 132102 ,2020

[11] Sana Belguith, “Enhancing Data Security in Cloud Computing Using a Lightweight Cryptographic Algorithm”, Conference: ICAS 2015: The Eleventh International Conference on Autonomic and Autonomous Systems

[12] K. Marimuthu, D. G. Gopal, K. S. Kanth, S. Setty and K. Tainwala, ”Scalable and secure data sharing for dynamic groups in cloud,” 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 2014, pp. 1697-1701.

[13] Sahai, A., Waters, B. “Fuzzy Identity-Based Encryption”. In: Cramer, R. (eds) Advances in Cryptology – EUROCRYPT 2005. Lecture Notes in Computer Science, vol 3494. Springer, Berlin.

[14] Chen, Y., Jiang, Z.L., Yiu, S.M., Liu, J.K., Au, M.H., Wang, X. (2015). “Fully Secure Ciphertext-Policy Attribute-Based Encryption with Security Mediator”. In: Hui, L., Qing, S., Shi, E., Yiu, S. (eds) Information and Communications Security. ICICS 2014. Lecture Notes in Computer Science(), vol 8958. Springer, Cham.

[15] C. -J. Wang and J. -F. Luo, ”A Key-policy Attribute-based Encryption Scheme with Constant Size Ciphertext,” 2012 Eighth International Conference on Computational Intelligence and Security, Guangzhou, China, 2012, pp. 447-451.

[16] H. Wang, S. Wu, M. Chen and W. Wang, ”Security protection between users and the mobile media cloud,” in IEEE Communications Magazine, vol. 52, no. 3, pp. 73-79, March 2014.

[17] S. K. Saroj, S. K. Chauhan, A. K. Sharma and S. Vats, ”Threshold Cryptography Based Data Security in Cloud Computing,” 2015 IEEE International Conference on Computational Intelligence and Communication Technology, Ghaziabad, India, 2015, pp. 202-207.

[18] K. -H. Lee and P. -L. Chiu, ”Digital Image Sharing by Diverse Image Media,” in IEEE Transactions on Information Forensics and Security, vol. 9, no. 1, pp. 88-98, Jan-2014.

[19] Mukherjee, R., and Ghoshal, N. (2013). Steganography Based Visual Cryptography (SBVC).

[20] Jon Henning “Exploring Real-World Cybersecurity Examples”, Coordinated Business Systems Blog, Aug 31, 2023.

[21] Murad, S.H., Rahouma, K.H. (2022).” Hybrid Cryptography for Cloud Security: Methodologies and Designs”. Digital Transformation Technology.

[22] Ms.Vaishnavi S. Kshirsagar1, Prof. N. M. Sawant, ” Privacy Protection for Cloud Based Online Transaction Using Steganography and Visual Cryptography”, International Journal of Innovations in Engineering and Science, Vol. 8, No. 5, 2023, PP. 42-44.

[23] Zadiraka, V.K., Kudin, A.M. “Cloud computing in cryptography and steganography”. Cybern Syst Anal 49, 584–588 (2013).

[24] Adee, Rose, and Haralambos Mouratidis. 2022. ”A Dynamic Four-Step Data Security Model for Data in Cloud Computing Based on Cryptography and Steganography” Sensors 22, no. 3: 1109.

[25] Devika, K.D., Saxena, A. “Dynamic Authentication Using Visual Cryptography”. In: Satapathy, S.C., Lin, J.CW., Wee, L.K., Bhateja, V., Rajesh, T.M. (eds) Computer Communication, Networking and IoT. Lecture Notes in Networks and

Systems, vol 459. Springer, Singapore.

[26] Palevicius, A., Janusas, G., Ragulskis, M., Palevicius, P., Sodah, A. (2018). “Design, Analysis, and Application of Dynamic Visual Cryptography for Visual Inspection of Biomedical Systems”. In Bonča, J., and Kruchinin, S. (eds), Nanostructured Materials for the Detection of CBRN. NATO Science for Peace and Security Series A: Chemistry and Biology. Springer, Dordrecht.

[27] Ch.RatnaBabu, M.Sridhar and Dr.B. RaveendraBabu, “Information hiding in greyscale images using pseudo-randomized visual cryptography algorithm for visual information security”, Proceedings of IEEE International Conference on Information Systems and Computer Networks (ISCON), at Gala University, Matura, ISBN:978-1-4673-5987-0, 9-10 March 2013, pp.195 – 199.

[28] Sultan Aldossary and William Allen, ”Data Security, Privacy, Availability and Integrity in Cloud Computing: Issues and Current Solutions”, International Journal of Advanced Computer Science and Applications(IJACSA), 7(4), 2016.

[29] Manjit Kaur, Ahmad Ali AlZubi, Dilbag Singh, Vijay Kumar, Heung-No Lee. ”Lightweight Biomedical Image Encryption Approach”, IEEE Access, 2023.

[30] Xuanxia Yao, Zhi Chen, Ye Tian, ”A lightweight attribute-based encryption scheme for the Internet of Things”, Future Generation Computer Systems, Volume 49, Pages 104-112, 2015.

[31] Ren, L., Zhang, D. “A QR code-based user-friendly visual cryptography scheme”. Sci Rep 12, 7667 (2022)

[32] V. Petrauskiene and L. Saunoriene, “Application of dynamic visual cryptography for optical control of chaotic oscillations,” Vibroengineering PROCEDIA, Vol. 15, pp. 81–87, Dec. 2017.

[33] Paulius Palevicius and Minvydas Ragulskis,” Image communication scheme based on dynamic visual cryptography and computer generated holography ”,Optics Communications, Volume 335, 2015, Pages 161-167.

[34] Charoghchi, S., Mashhadi, S. “A secure secret image sharing with steganography and authentication by Hamming code (15,11) for compressed images”. Multimed Tools Appl 83, 31933–31955 (2024).

[35] Ahmad, S., Abidi, M.R. (2022). “RGB Based Secure Share Creation in Steganography with ECC and DNN”. In: Unhelker, B., Pandey, H.M., Raj, G. (eds) Applications of Artificial Intelligence and Machine Learning. Lecture Notes in Electrical Engineering, vol 925. Springer, Singapore.

[36] Salim, A., Sagheer, A.M., Yaseen, L. (2020). “Design and Implementation of a Secure Mobile Banking System Based on Elliptic Curve Integrated Encryption Schema”. In: Khalaf, M., Al-Jumeily, D., Lisitsa, A. (eds) Applied Computing to Support Industry: Innovation and Technology. ACRIT 2019. Communications in Computer and Information Science, vol 1174. Springer

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Mar 23, 2024
Accepted: Jun 22, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 96–113, DOI 10.12694/scpe.v26i1.3826

IOT ENABLED SMART AGRICULTURE SYSTEM FOR DETECTION AND CLASSIFICATION OF TOMATO AND BRINJAL PLANT LEAVES DISEASE

ROHIT KUMAR KASERA [*], SWARNALI NATH, BIKASH DAS, ANIKET KUMAR, AND TAPODHIR ACHARJEE

Abstract. Internet of Things (IoT) assisted smart farming techniques are gradually being used efficiently for identification and classification of vegetable plant diseases. Detection and classification of diseases in these plant families like Solanaceae are still problematic using DCNN due to variations in environmental conditions, genome variation, type of disease, etc. In this paper, two methods for spotting and diagnosing diseases of brinjal and tomato plants leaves named as Optimal Environmental Traversing Alert (OETA) and Optimum diagnosis of Solanaceae leaf diseases (ODSLD) respectively have been proposed. The OETA machine learning (ML) based method is used first to detect the disease, and then the ODSLD deep convolutional neural networks (DCNN) method is used to classify it. An analysis of the proposed method experiments showed that OETA disease detection for brinjal plant (eggplants) was 97.81 percent and for tomato plants was 99.03 percent. For disease classification by ODSLD method, the VGG-16 for brinjal plant and ResNet-50 for tomato plants outperformed other existing DCNN computer vision methods.

Key words: Smart farming, Disease detection, VGG19, DenseNet121, Edge computing, Raspberry pi pico

1. Introduction. India and several other nations rely substantially on agriculture as their primary source of income. It has an enormous influence on the economic development of many nations [1]. In traditional farming, farmers in remote areas face obstacles in following and assessing meteorological conditions, soil quality, water availability, insect control, disease identification, regular monitoring of farming field, and other factors [2]. The majority of leaflets get struck with multiple diseases before and during harvest, which diminishes crop quality and yields. Every year a lot of plants and crops are destroyed due to various causes, including fungal microorganisms pH imbalances in the soil, severe temperatures, alterations in atmospheric moisture or humidity, an inadequate volume of nutrients in the soil, and other aspects [3]. This work of disease detection research focuses on two species of plants from the nightshade family (Solanaceae) [4] namely tomato and brinjal (Eggplant). The diseases that mostly affect tomato and brinjal crops are given in Fig. 1.1 and 1.2.

We have selected tomato and brinjal in our study because both are cultivated simultaneously by farmers across India in dry seasons and and are highly susceptible to diseases [5]. So if a single system can be used to predict and classify disease in the plants, it will be quite a cheap software for farmers. Using precision farming technology with the Internet of Things (IoT), farmers can effortlessly recognise the category of diseases that harms a leaf of tomato and brinjal, thereby improving yield and production [6]. The automated identification and categorization of diseases is an increasingly prominent area of research in smart agriculture today. Various of research have been been pulled off for the recognition and categorization of crop illnesses by employing predictive modelling or Deep convolutional neural networks (DCNN) approaches involving Support vector machines (SVM), Random forest (RF) [7], Artificial neural networks (ANN) [8], Decision trees (DT) [9], Convolutional neural networks (CNN), Visual Geometry Group-16 (VGG16), Inception v3, DenseNet-121, and Residual Network 50 (ResNet 50), U-Net [10, 11]. Sophisticated smart farming systems could be enhanced by combining IoT, machine learning (ML), and DCNN methods. Precisely anticipating and classifying crop health metrics for boosting yield in agriculture is one of the main obstacles. Due to lack of accurate, proficient data and efficient predictive models, genetic factors of plants, and variability in leaf images it is challenging to develop a real-time decision system. Farmers often have difficulty in determining well-informed decisions concerning crop monitoring, pest control, and irrigation without real-time disease detection systems, involving soil nutrient data and a breakdown of tomato and brinjal crop disease types. As a consequence, reliable and

[*] Department of Computer Science & Engineering, Triguna Sen School of Technology, Assam University, Silchar, Assam, India. (rohitkumar.kasera@aus.ac.in)

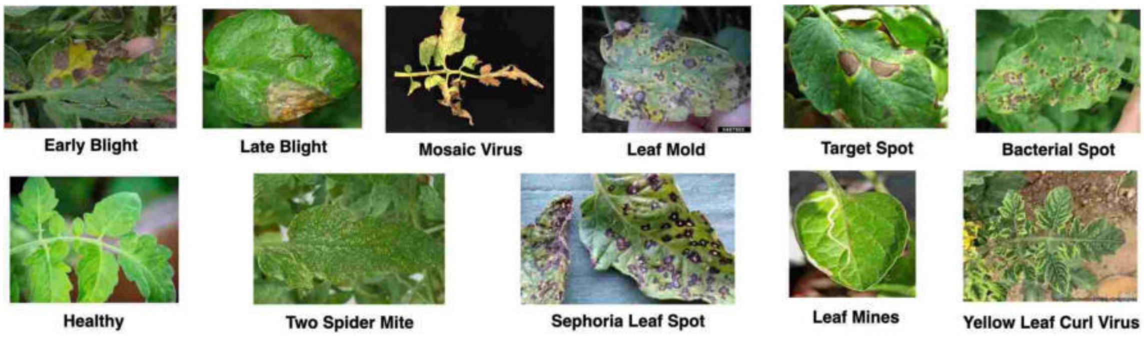


Fig. 1.1: Segmentation of tomato ailments into classes

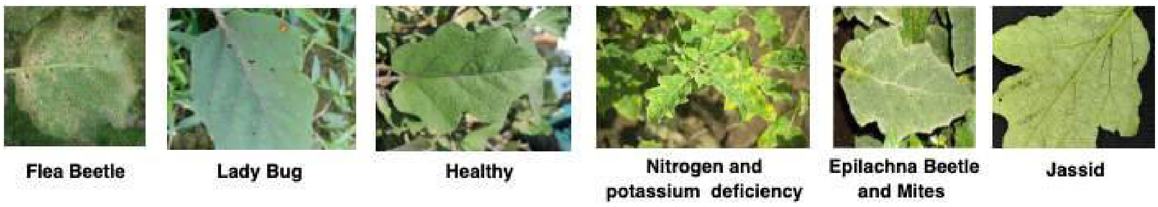


Fig. 1.2: Segmentation of brinjal plant ailments into classes

Table 1.1: Optimal Ranges values for Tomato and brinjal plant cultivation

Plant	**Temp** [$^{\circ}C$]	**Hum** [%]	**SM** [%]	**pH**	**N** [mg/Kg]	**P** [mg/Kg]	**K** [mg/Kg]
Tomato	10 -28	60 - 80	60 - 90	5 - 8	40 - 120	30 - 110	30 - 100
Brinjal	20 - 28	60 - 80	65 - 90	4 - 7	60 - 120	50 - 80	50 - 90

scalable hybrid Edge IoT and AI-enabled alternatives have to exist for the ability to effectively detect illnesses, forecast crop conditions, and allocate resources in intelligent agriculture systems [26]. Recent research indicates the optimal temperature (Temp), humidity (Hum), soil moisture (SM), pH, nitrogen (N) phosphorus (P) and potassium (K) values ranges for growing cultivating tomatoes [12, 13] and brinjal [14] are shown in the Table 1.1. N, P, and K scales on the soil can be used to gauge tomato and Brinjal plant growth. It is possible to avoid both inadequate nutrient supply and over fertilization by anticipating nutrient data. Consequently, farmers can make informed decisions based on real-time temperature, humidity, soil moisture, pH, N, P, K levels and determine the exact amount of fertilizer required for robust plant growth [15], and nutrient level information will help the ML methods to decide the optimal disease prediction.

The intended effect of this research is to assist low-income farmers by predicting and recognizing types of diseases in the early stage through the adoption of an amended IoT edge AI-based methods as Optimal Environmental Traversing Alert (OETA) and Optimum diagnosis of Solanaceae leaf diseases (ODSLD). The proposed system can be employed with some initial cost (specially season cost) in first years but following years only minimal maintenance costs are to be borne by the farmers. This research's key contributions are outlined below:

1. Real-time data through sensor including camera were collected from tomato and brinjal plants fields using an LPWAN communication method. These data are used to develop a disease monitoring system that can detect and recognize diseases.
2. This article adopts an innovative approach utilizing an IoT Edge-based disease observation module called the OETA prognosis method. It employs a modified stacked ensemble learning method. This

method collects data from the surrounding environment (temperature and humidity) and soil (Moisture, pH, N, P, and K) of tomato and brinjal growing fields in real-time to predict the illness and soil health status.

3. The ODSLD hybrid DCNN computer vision technique was designed to classify the disease found on brinjal and tomato plants.
4. An extensive comparative analysis of the hybrid disease testbed framework has been accomplished through experimental results and discussions to validate its performance.

This article is structured into several sections, the Sect. 1 describes the problem domain, possible discrepancies in existing research, objectives, and contributions to strengthen the IoT-enabled illness detection system. Sect. 2 covers in-depth literature surveys along with their limitations and research deficiencies; Sect. 3 demonstrates the proposed methodology; Sect. 4 unveils experiment results and an assessment of the proposed methodology with comparative analysis; and Sect. 5 conveys a summary of the proposed system with future research.

2. Literature Review. In [16] researchers proposed an IoT monitoring system to analyze the minimal and extreme ranges of every environmental factor to envision four types of plant diseases. For classification, they used ANNs and achieved more than 98% accuracy, but the system is not fully automated and adaptive and has very few sensors.

A disease recognition framework for tomato and brinjal plant has been put forward using SE-Inception in [18]. The model utilizes a multi-scale mining module to boost its efficiency and a batch normalization layer to accelerate network convergence. 98.29% accuracy was accomplished with this particular strategy. The assessment of this methodology's flaws indicates floating point arithmetic can impact the mobile device's efficiency although real-time data hasn't been utilized during any kind of system validation.

In [17] the authors put forward a two-stream classification strategy using Inception V3 and inference with recognition using ”CNN-softmax”. It is observed that the system is not fully automated and precision is low.

Research in [19] indicates utilizing OPNN-based plant disease diagnosis to identify brinjal leaf problems early on. RGB transformation, pre-processing with a median filter, feature extraction with ideal weight values, and ARMKFCM segmentation of impacted areas are all-encompassed. The main flaw of the prevailing approach is its inadequate datasets.

In the article [21], an apparatus is utilized to grab an image and transmit it to a neural network for categorization. Considering brinjal plant of sixteen classes, they adopted CNN transfer learning (”DenseNet201, Xception, ResNet152V2”). DenseNet 201's effectiveness reached 99.06%. The suggested approach has a few limitations, such as skipping the dataset scale constraint. Transfer learning causes an overfitting problem during validation.

Tomato plants are allowed to acquire environmental data with the assistance of an IoT module in [20]. Models like Random Forest, SVM, and K-means are used for assessing the health of plants for three sickness classes and one healthy class. Vanilla Architecture, VGG16, and VGG19 were used to gauge how well the suggested models performed. VGG16 obtained a 92.08% accuracy rate. The outcomes can be upgraded by collaborative prediction, and the efficacy of the model is not validated.

Incorporating IoT and machine learning strategies in [22], IQWO-PCA is an enhanced quantum whale optimization practice that foresees plant epidemics in farming. The system has a few limitations, such as a lack of information on the type of IoT devices used to set up the experiment testbed. This model cannot classify many types of tomato diseases because the system has been tested on a single disease type.

In [23], a CNN-modified imitation has been conceived to predict tomato leaf diseases from a 50k dataset of 14 crops with a performance at 91.2%. In its entirety, CNN triumphed over other pre-trained models, notably VGG16, InceptionV3, and MobileNet, in contrast to their contributions. A few limitations are the inadequate performance and deficient adaptability of the suggested model.

The aforementioned survey emphasizes the ongoing research gap concerning the performance involving various computer vision-based illness detection methods, gradient vanishing, system responsiveness, recognizing between healthy and damaged plants, adoption of IoT devices for real time system, Long range data transmission, sensor data processing, and automated computational tasks.

Table 3.1: Unprocessed sensor data

Time	Temp	Hum	SM	pH	N	P	K	Light
14/12/2023 10:30:00 AM	20.23	74	62	6.05	120	61	92	278
08/02/2024 7:30:00 PM	15.23	73	81	4.23	112	56	87	112
13/03/2024 12:45:00 PM	30.8	69	52	5.62	117	47	67	338
13/03/2024 01:30:00 PM	30.01	72	57	4.02	117	47	67	356

3. Methodology. Considering the research gaps which are addressed in Sect. 2, a hybrid IoT edge-enabled disease diagnosis approach is put forward in this current work. The two methods compose the overall testbed of the system. The first predicts disease using the Optimum Environmental Traversing Alert (OETA) method based on IoT edge and ML, and the second classifies types of diseases using image processing and DCNN named as optimal diagnosis of Solanaceae leaf diseases (ODSLD). The comprehensive architecture for automated disease detection in tomatoes and brinjal plant is portrayed in Fig. 3.1.

The disease recognition framework in Fig. 3.1 is an Internet of Things (IoT) edge-based architecture composed of sensing layer unit (SLU), IoT edge gateway unit (IGU), cloud and application layer unit (ALU). The SLU processes real-time sensor data on temperature (Temp), humidity (Hum), soil moisture (SM), pH, nitrogen (N), phosphorus (P), potassium (K), and light through Raspberry Pi Pico WH microcontroller from an open tomato and brinjal field. Sensor data is transmitted in real-time via the LoRa SX1278 433Mhz module to an IGU. Real-time transmission over a low power wide area network (LPWAN) at long range establishes a point-to-point (P2P) connection between SLU and IGU. The Raspberry Pi 4 module is used as an IoT edge gateway (IGU) for entire system testing. The accumulated data from SLU is processed under IGU. The data are preprocessed and then trained using the OETA classification method, which is a modified hybrid machine learning (ML) ensemble classifier. Hybrid ensemble approaches assess the abilities of more than one robust ML model in a classification task. It reduces overfitting, maximizes true positives, and contributes to an imbalanced dataset. An "OV7670" camera module is deployed under the SLU. If the OETA method predicts that the tomato or brinjal crop has a disease based on the current state of the real-time sensor data, IGU will instruct SLU to activate the camera to take a picture of the leaf. SLU will transmit the captured image to IGU via LoRa SX1278 to recognize the type of disease. The transmission of image data through the LoRa module is accomplished based on the existing multi-packet LoRa transmission protocol [24] using a lightweight Joint Photographic Experts Group (JPEG) coder [25]. The accumulated captured leaf images are stored in the local storage of the Raspberry Pi 4 to recognize the type of leaf disease that occurred. The stored leaf images are preprocessed to remove outliers through image processing. Following preprocessing, these leaf image data are trained using a hybrid deep convolutional neural networks (DCNN) model named ODSLD for tomato and brinjal disease classification under IGU. Once the ODSLD model has been learned, it is tested and validated. Validation involves evaluating the predicted disease classification, which recognizes the particular tomato or brinjal leaf disease name. Farmers Solanaceae disease dashboard provides daily, next-day, and disease prediction alerts which can be accessed through mobile or laptop.

3.1. IoT edge based disease detection system. As shown in Fig. 3.2 (a), a LoRa enabled SLU prototype setup is placed near tomato and brinjal plants to accumulate real-time sensor data, capture images, and process them. Fig. 3.2 (b) shows the LoRa-enabled IoT edge gateway (IGU) setup which is used to receive the tomato and brinjal crop sensor data for further processing.

The sensors are mounted in an array to grab the data. For the day, the sensor reveals the data values at three-minute intervals. Eight traits have been acquired for tomatoes: Temperature (Ft_1), Humidity (Ft_2), Soil moisture (Ft_3), pH (Ft_4), N (Ft_5), P (Ft_6), K (Ft_7), and light (Ft_8). Regarding brinjal, the eight amenities that are collected are Temperature (Fb_1), Humidity (Fb_2), soil moisture (Fb_3), pH (Fb_4), N (Fb_5), P (Fb_6), K (Fb_7); and light (Fb_8). The sensor data in .csv format are extracted to create the dataset for IoT edge-based illness training and prediction. It is a three-minute-long real-time dataset with various attributes. A total of 20532 entries of data accumulated for tomato and 10236 for brinjal plant between 20-11-2023 to 25-03-2024. Table. 3.1 shows raw data of both crops based on sensor readings.

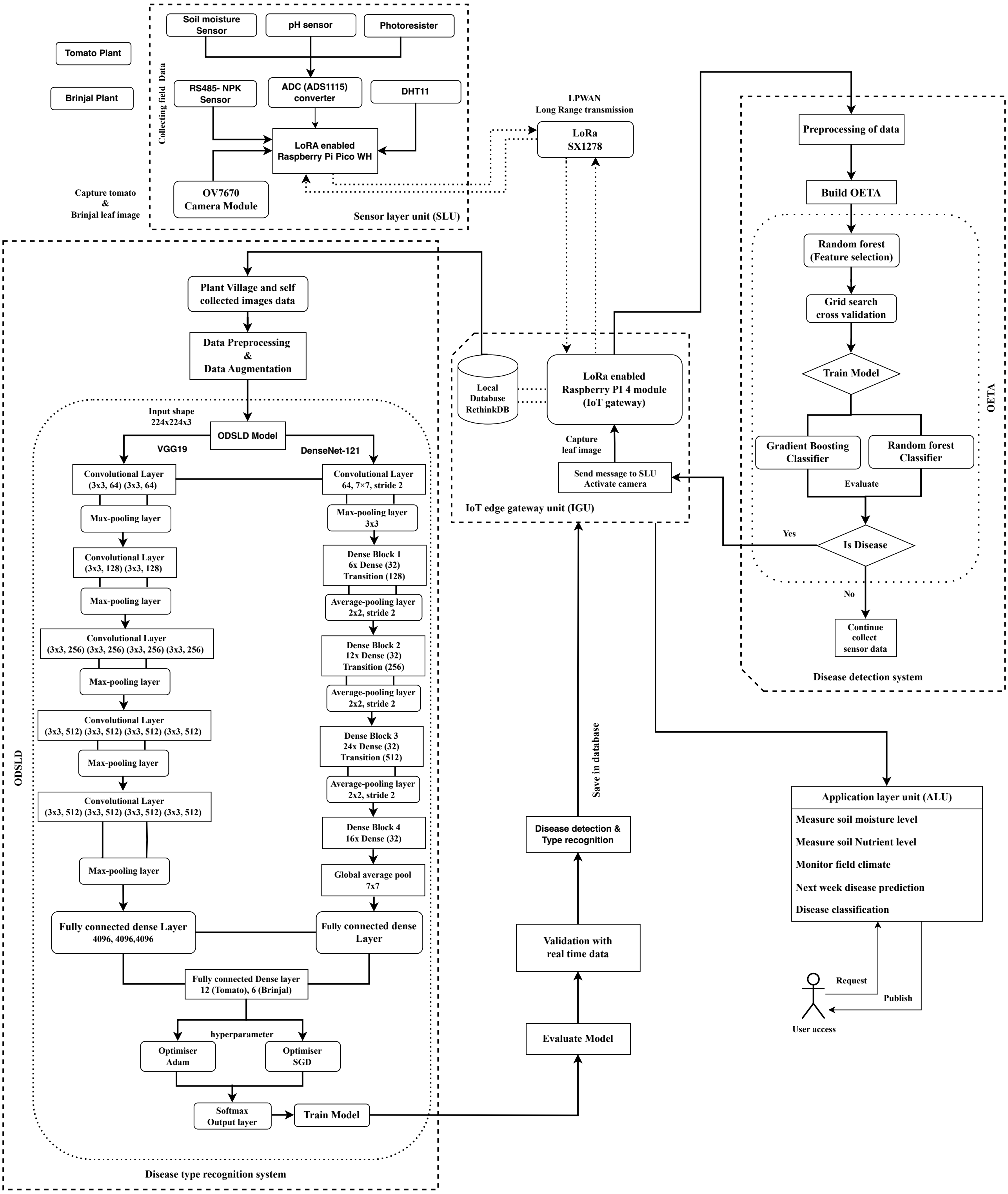

Tomato Plant
Brinjal Plant
Soil moisture Sensor
pH sensor
Photoresister
Collecting field Data
RS485- NPK Sensor
ADC (ADS1115) converter
DHT11
LoRa enabled Raspberry Pi Pico WH
Capture tomato & Brinjal leaf image
OV7670 Camera Module
Sensor layer unit (SLU)
LPWAN Long Range transmission
LoRa SX1278
Plant Village and self collected images data
Data Preprocessing & Data Augmentation
Input shape 224x224x3
ODSLD Model
VGG19
DenseNet-121
Convolutional Layer (3x3, 64) (3x3, 64)
Max-pooling layer
Convolutional Layer (3x3, 128) (3x3, 128)
Max-pooling layer
Convolutional Layer (3x3, 256) (3x3, 256) (3x3, 256) (3x3, 256)
Max-pooling layer
Convolutional Layer (3x3, 512) (3x3, 512) (3x3, 512) (3x3, 512)
Max-pooling layer
Convolutional Layer (3x3, 512) (3x3, 512) (3x3, 512) (3x3, 512)
Max-pooling layer
Fully connected dense Layer 4096, 4096,4096
Convolutional Layer 64, 7×7, stride 2
Max-pooling layer 3x3
Dense Block 1 6x Dense (32) Transition (128)
Average-pooling layer 2x2, stride 2
Dense Block 2 12x Dense (32) Transition (256)
Average-pooling layer 2x2, stride 2
Dense Block 3 24x Dense (32) Transition (512)
Average-pooling layer 2x2, stride 2
Dense Block 4 16x Dense (32)
Global average pool 7x7
Fully connected dense Layer
Fully connected Dense layer 12 (Tomato), 6 (Brinjal)
hyperparameter
Optimiser Adam
Optimiser SGD
Softmax Output layer
Train Model
ODSLD
Disease type recognition system
Local Database RethinkDB
LoRa enabled Raspberry PI 4 module (IoT gateway)
Capture leaf image
Send message to SLU Activate camera
IoT edge gateway unit (IGU)
Save in database
Disease detection & Type recognition
Validation with real time data
Evaluate Model
Preprocessing of data
Build OETA
Random forest (Feature selection)
Grid search cross validation
Train Model
Gradient Boosting Classifier
Random forest Classifier
Evaluate
Yes
Is Disease
No
Continue collect sensor data
OETA
Disease detection system
Application layer unit (ALU)
Measure soil moisture level
Measure soil Nutrient level
Monitor field climate
Next week disease prediction
Disease classification
Request
Publish
User access

Fig. 3.1: Framework for IoT-based disease detection and classification

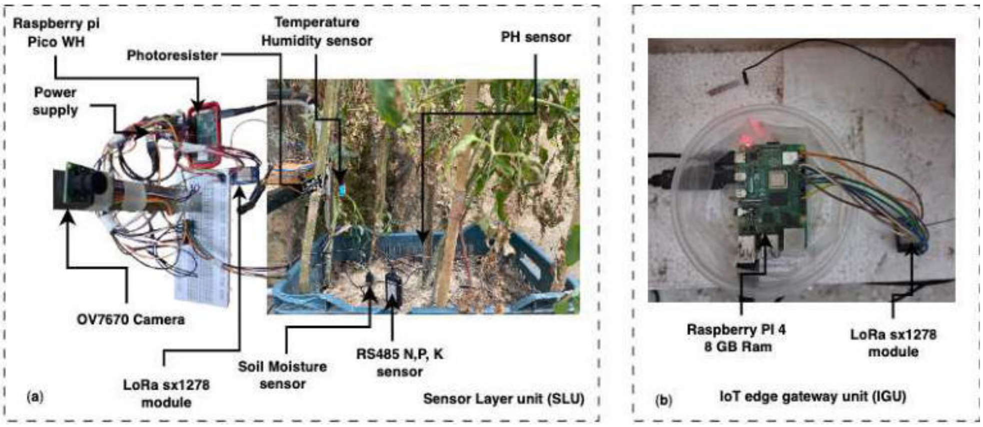


Fig. 3.2: Prototype setup of IoT Edge based disease detection module

The unprocessed data are first filtered by labeling, normalizing, and wiping missing and negative values. The data have been normalized by incorporating five minutes for all features. As discussed in Sect. 1, the optimal value range for each feature has been considered to set the label data $t_{response}$ as 0 or 1. A label data row with '1' indicates that any of the features is outside the range (diseased), while '0' indicates that it is in between or equal the range (Non disease). OETA ML is a converged method of Random Forests (RF), gradient boosting (GB), and stacked classifiers. It is used as a base model for training OETA. This hybrid method aims at filling the shortcomings of individual ML methods. As a result, overfitting is reduced and large and small datasets are handled efficiently. In this necessary features are selected based on predicted RF feature scores and make predictions through grid search and parameter optimization of each model. As part of the process of building the OETA method, the following steps are listed with mathematical representations.

- A Z-score normalization is achieved using Eq. (3.1). This method standardizes data by scaling fx_i features.

$$Zx_i \rightarrow \frac{fx_i - M\mu_i}{S\sigma_i} \tag{3.1}$$

In the Eq. (3.1), Zx_i is the normalized feature, $M\mu_i$ is the intermediary, and $S\sigma_i$ is the standard error of the fx_i feature.

- An RF classifier is used to estimate important features score for the feature fx_i in Eq. (3.2).

$$pivotal(fx_i) \geq threshold \tag{3.2}$$

In Eq. (3.2), the threshold is the average of all features' importance scores, and pivotal (fx_i) is expressed as importance. Table. 3.2 summarizes the statistical evaluation of the real-time sensor data after normalization and feature extraction.

- The base RF and GB model is converged by assembling to build and train the OETA model.

$$\hat{O}_{mf} \rightarrow \frac{1}{K}\sum_{l=1}^{K} H_l(fx) \tag{3.3}$$

In the Eq. (3.3), a random decision tree is concatenated to enhance predictive performance. Where prediction $\hat{O}_{mf}$ is the mean prediction of K trees, and $H_l(fx)$ is the prediction of l^{th} tree.

$$\hat{O}_{gb} \rightarrow \sum_{k=1}^{K} n * w_k(fx) \tag{3.4}$$

Table 3.2: Statistical summary of feature variable

Feature variable	**Mean**	**Standard Deviation**
Temperature	28.60	1.76
Humidity	85.29	12.30
Soil moisture	314.24	74.58
pH	3.34	0.40
N	117.70	37.24
P	42.86	25.44
K	60.0	27.72

A tree is formed iteratively using GB classifier to optimize prediction performance and minimize error. In Eq. (3.4), $\hat{O}_{gb}$ describes the final prediction of the model, n depicts the learning rate, $w_k(fx)$ defines the prediction of the k_{th} iteration, and K expresses frequency of boasting loops.

$$\hat{O}_{converge} \to U_{ultimate}([\hat{O}_{mf}, \hat{O}_{gb}]) \tag{3.5}$$

Eq. (3.5) shows base model is converged after evaluating and its prediction is used as a feature to build and train the OETA classifier model, where OETA ultimate prediction output as $\hat{O}_{converge}$, and $U_{ultimate}$ is the OETA classifier which incorporates all of the predictions.

- Each parameter in the base model and OETA classifier is optimized by grid search cross-validation using Eq. (3.6).

$$score_{hyperparameter} \to crossVal_{score}(U_{converge}, S_{feature}, t_{response}) \tag{3.6}$$

In Eq. (3.6), $S_{feature}$ is the selected features evaluated using Eq. (3.2) can contain the arrays of feature as Ft_1, Ft_2,...,Ft_8 for tomatoes and Fb_1, Fb_2,...,Ft_8 for brinjal plant, $t_{response}$ denotes the target data which contains two classes, and $U_{converge}$ denotes the ultimate ensemble classifier.

- The soil nutrient recommendation is based on current nutrient levels and specified thresholds.

$$\text{If } fx_i < \text{limit}_{\text{low}}, \text{then Increase } fx_i$$

$$\text{If } fx_i > \text{limit}_{\text{high}}, \text{then Decrease } fx_i$$

Several evaluation metrics, namely mean squared error (MSE), mean absolute error (MAE), and root mean square error (RMSE), seemed applied to assess the efficiency of the OETA model. Furthermore, the efficiency of the suggested approach was also validated using additional ML classifiers, notably Support vector machine (SVM), Logistic regression (LR), Decision tree (DT), and RF.

3.2. Disease classification and recognition system. The plant leaf diseases are recognised using the hybrid computer vision deep convolutional neural networks (DCNN) approach named as Optimum diagnosis of Solanaceae leaf diseases (ODSLD) for tomato and brinjal plants. Existing CNN-based models for tomato and leaf disease classification have various challenges and shortcomings. These include Precise feature extraction, effectiveness of model learning, diminishing gradients, intricate datasets, reliability of the model, overfitting challenges, dense connectivity, and maximal illness classification. To overcome these challenges, the ODSLD model improves in-depth and reusable feature extraction from images. It strengthens gradient flow, which means there is less chance of vanishing gradients, and efficiently manages small and complex datasets. Overfitting is less probable as features are reused with fewer parameters, maximizing the prediction performance and model learning efficiency using multiple optimizers. The ODSLD DCNN model is a hybrid convolutional neural network (CNN) derived from VGG19 [27] and DenseNet121 [28]. The ODSLD model architecture is depicted in Fig. 3.1 for categorizing tomato and brinjal disease types. The model consists of 137 convolutional layers, 9 pooling layers, global pooling layer, 4 fully connected dense layers, and 1 output softmax layer. Data processing and training of ODSLD models are explained in the following steps.

- Kaggle - Tomato Leaves Dataset [29] is the origin of the dataset for the tomato plant, whereas Kaggle - Brinjal Diseases Detection [30] is the origin of the dataset for the brinjal plant.
- The tomato plant dataset contains 12 classes of images, with 5024 images and 95 megabyte (MB) data size, representing 11 types of diseases. The diseases are Early Blight, Late Blight, Septoria Leaf Spot, Tomato Yellow Leaf Curl Virus, Bacterial Spot, Target Spot, Leaf Mold, Tomato Mosaic Virus, Spider mites Two-spotted, Powdery Mildew and a new disease class, Leaf Mines along with a healthy class. This Leaf Mines disease is caused by the larvae of Tuta absoluta [31] that causes white splotches on the tomato leaves.
- The brinjal plant dataset contains 1235 images of 115 MB data size consisting of six classes. Among the six classes, one is healthy, and the other five are diseased: Epilachna Beetle and Mites, Flea Beetle, Jassid, Potassium, and Nitrogen Deficiency, and one new species is included known as Ladybugs.
- In furtherance of pre-existing datasets, real-time datasets were additionally gathered through visits to outdoor tomato and brinjal agricultural fields in Assam, India, to test and validate the system.
- Preliminary processing entails working on the gathered image data to eradicate undesired images (blurred and out-of-frame images). It entails scaling the photos of the brinjal and tomato leaves to 256x256. Flip and rotation at 90, 180, and 270 degrees were additionally applied to raise the amount of images and zooming in or out is a possibility up to 20%. 10% of the dataset has been split for the validation set. Once the data is adequately collected and preprocessed, it is fed into the ODSLD model for training.
- We used two dense layer hyperparameters. First, dense layer hyperparameters have a minimum value of 128 and a maximum value of 512, the second has a minimum value of 64 and a maximum value of 256 with a RELU activation function.

The custom random search method configures the Keras tuner and optimizes the model hyperparameter using training and validation data. Data have been encapsulated in a custom method to assemble images and labels. In the ODSLD model tuner search, an optimizer is automatically selected during training between ADAM and SGD, and the learning rate is altered correspondingly. In this way, it diminishes the catastrophic over-fitting issue in image-driven plant diagnosis tasks and improve the ODSLD model accuracy. A hyperparameter tuning process determines the best model and evaluates its performance. The reliability of the ODSLD model has been validated using metric factors, which include reliability using Eq. (3.7), true positive rate (TPR) using Eq. (3.8), precision using Eq. (3.9), and F1-score is measured using Eq. (3.10) [32]. Were AO is the actual optimistic, AU is actual unfavourable, DO is deceptive optimistic, and DU is deceptive unfavourable.

$$reliability \rightarrow \frac{AO + AU}{AO + AU + DO + DU} \tag{3.7}$$

$$TPR \rightarrow \frac{AO}{AO + DO} \tag{3.8}$$

$$precision \rightarrow \frac{AO}{AO + DU} \tag{3.9}$$

$$F1score \rightarrow \frac{2 * TPR * precision}{TPR * precision} \tag{3.10}$$

A Raspberry Pi pico camera captures the image of the tomato or brinjal plant leaf and transmits it to the IoT edge gateway in real-time for further processing. The received leaf images are preprocessed into 224x224 size and entered into the ODSLD model to predict the disease class. The predicted disease class will be stored in the real-time rethinkDB database to access. IoT edge-based diagnosis system uses a lightweight message-queued telemetry transport protocol (MQTT) to transmit data. When the real-time disease system receives subscribed disease notifications, it will publish them to ALU through the MQTT broker. Overall, the system is automated and runs computations in real time. The proposed intelligence disease framework assists farmers in preventing infections in tomato and brinjal farms, hence increasing the production of brinjal and tomatoes.

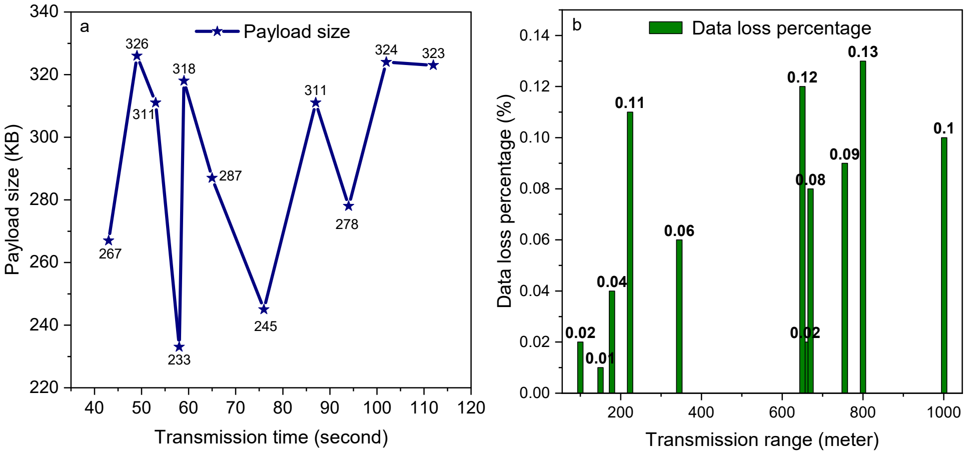


Fig. 4.1: Tomato and brinjal field transmission measurement between SLU to IGU

4. Experiment Results and Discussion. In the Raspberry Pi IoT edge environment, overall system frameworks have been built in Arduino, C++ and Python. Fig. 4.1 (a) and (b) shows the long range wide area network (LoRaWAN) transmission duty cycle of tomato and brinjal crop sensor data. Fig. 4.1 (a) shows the sensor data transmission analysis between SLU and IGU based on payload size (KB) and transmission time (seconds). Fig. 4.1 (b) shows the analysis of data loss percentage based on transmission range. It can be seen from Fig. 4.1 that LoRaWAN transmission between SLU and IGU performs better in terms of payload size, long-range transmission, and lower loss rate. The average transmission time takes up to 61 seconds to transmit the sensor data with a 0.17% loss rate from SLU to IGU.

4.1. Disease detection performance analysis. Real-time tomato and brinjal crop data collected from the open field are normalized as discussed in 3.1. A random forest (RF) ML method was applied to identify essential feature variables. A heatmap matrix shown in Fig. 4.2 depicts the correlation between the feature variables. Fig. 4.2 correlation matrix demonstrates how favorably all feature variables correlate with each other. There is a strong correlation between the variables temperature, humidity, soil moisture, and potassium. There is a positive correlation among the variables potassium, phosphorus, nitrogen, and pH. The strongest correlation is between nitrogen and phosphorus variables.

OETA ML model was trained on 70% training data and 30% testing data. Across the training phase of OETA models of tomato and brinjal crops, a K-split cross-validation at a 5-fold cross is used from which it optimized the mean value, alleviated bias and inconsistency, balanced the data and enhanced data utilization owing to every data point utilized to assess its model's efficiency. The OETA model parameters are the gradient boosting (GB) and random forest (RF) convergence hyperparameters, configured as OETA model-tuning hyperparameters. The optimal hyperparameters on which OETA model performance improves are gb___learning_rate 0.1, gb___max_depth 3, gb___n_estimators 100, rf___min_samples_leaf 1, rf___min_samples_split 2, and rf___n_estimators 100. Using the best hyperparameters, the OETA model achieved 98.86% testing accuracy for brinjal plant and 99.23% accuracy for tomato plant. The ROC of the OETA model has been evaluated as 0.97 for the brinjal plant and 0.99 for the tomato plant. Fig. 4.3 (a) and (b) show a plotted graph of the ROC curve based on the X-axis as the False positive rate and the Y-axis as the True positive rate.

To validate the OETA ML model using real-time sensor data, the model has been dumped using Joblib. Table. 4.1 and 4.2 display the overall validation classification report of both crop from various ML and OETA method for disease detection using real time sensor data.

The classification score analysis revealed in Table. 4.1 and 4.2 that the OETA ML model edges out another ML model in terms of disease prediction for tomato and brinjal plants. The estimations show that tomato plants had an RMSE of 0.08 and brinjal plants had an RMSE of 0.12 in disease prediction. Table. 4.3 shows

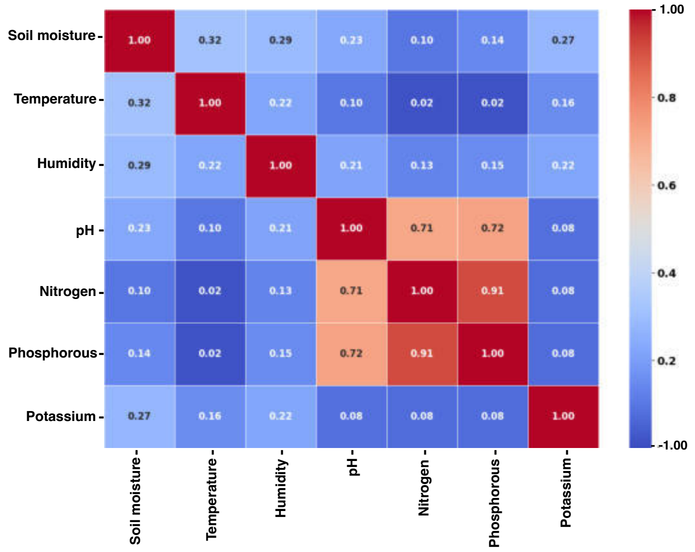


Fig. 4.2: Correlation coefficients matrices between the set of selected features variable

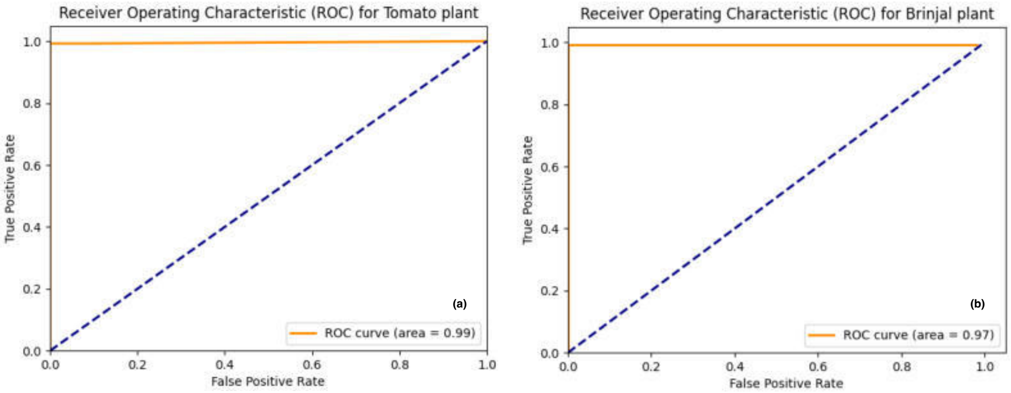


Fig. 4.3: OETA model ROC curve analysis for Tomato and Brinjal plant disease detection

Table 4.1: OETA and other ML model performance report for brinjal plant disease detection

ML Model	Accuracy	F1-score	Precision	Recall	MSE	MAE
OETA	98.56	98.90	98.50	97.90	0.09	0.08
SVM	88.12	89.61	90.13	89.11	3.78	3.23
RF	95.21	95.55	95.16	95.97	3.56	3.65
DT	94.07	94.55	95.16	94.97	2.88	2.53
LR	88.75	89.18	90.12	88.26	4.34	3.67

the disease prediction performance of the OETA method with other existing ML methods through real-time data.

Table. 4.3 shows that whenever the climate conditions, soil moisture, and nutrients are in the range of disease (1) or not disease (0), the proposed OETA method performs better prediction than the other existing

Table 4.2: OETA and other ML model performance report for tomato plant disease detection

ML Model	Accuracy	F1-score	Precision	Recall	MSE	MAE
OETA	99.07	99.16	99.32	99.11	0.07	0.08
SVM	87.15	86.17	86.71	86.59	4.12	3.43
RF	94.26	95.49	95.56	95.18	2.78	2.59
DT	93.15	94.21	95.11	94.35	3.14	3.08
LR	85.47	88.13	85.62	89.16	5.21	5.28

Table 4.3: Comparative prediction performance analysis of OETA with existing ML technique

ML Model	Temp	Hum	SM	N	P	K	Actual	Predicted
OETA	29.80	61.70	113	134	97	76	1	1
SVM	31.32	86.17	634	124	66	83	1	0
RF	30.04	78.17	546	115	123	118	1	0
DT	20.04	85.17	213	103	90	95	0	1
LR	21.14	67.17	278	111	91	94	0	1

Table 4.4: Summary of GPU processor used for training ODSLD method

Plant leaf	Memory usage	Average GPU usage	Total training time
Tomato	70230 MB	48%	12 hours
Brinjal	65216 MB	36%	3 hours

Table 4.5: Comparative performance analysis of DCNN classification model for tomato plant

Method	Test Loss	Validation Loss	Learning rate	Optimizer
ODSLD	0.032	0.017	0.0098	SGD
ResNet-50	2.41	2.36	0.0045	Adam
Inception-V3	3.13	3.22	0.0098	Adam
VGG16	2.11	1.88	0.0034	Adam
CNN	5.26	5.38	0.0098	Adam
VGG19+ResNet-50	1.26	1.38	0.0076	Adam

machine learning methods. When analyzing the prediction performance of conventional RF methods, we observe that the prediction value suggests no disease, and the actual value suggests a disease. The RF method accuracy for tomato and brinjal plants is around 94%, as shown in Table. 4.1 and 4.2.

4.2. Disease classification method performance analysis. Optimum diagnosis of Solanaceae leaf diseases (ODSLD) using deep neural networks model is trained using TensorFlow 2.17, Keras tuner random search, and NUMPY. Data is split into 80% training, 10% testing and 10% validation using Keras' preprocessing module to introduce randomness and shuffle. The model has been trained at 50 epochs throughout each trial using various combinations of hyperparameters for a maximum of 10 trials. This approach optimizes model learning by training the model on each hyperparameter with batch sizes of 16 respectively. The ODSLD model's total parameter size is 176.87 megabytes (MB), the trainable parameter size is 176.55 MB, and the non-trainable parameter is 326.75 kilobytes (KB). A summary of Graphics processing unit (GPU) resources used for constructing and training the ODSLD DCNN model is shown in Table. 4.4. The implementation has been done in a GPU with 81920 MB memory and 70350 MB processing capacity in 12.08 hours. The performance of ODSLD models with existing DCNN models for the tomato plant is validated in Fig. 4.4 and Table. 4.5.

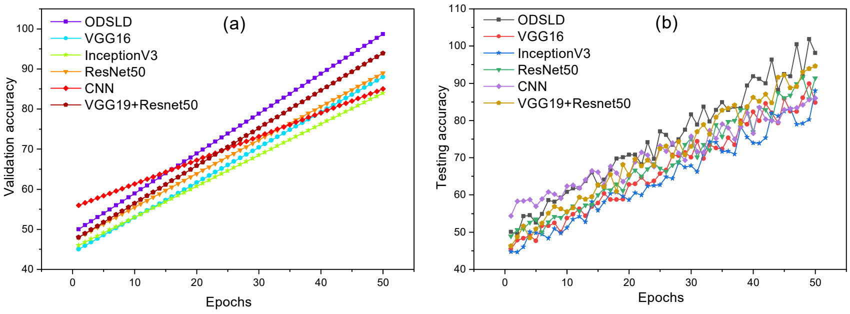


Fig. 4.4: Testing and validation performance curve for tomato plant using real time data

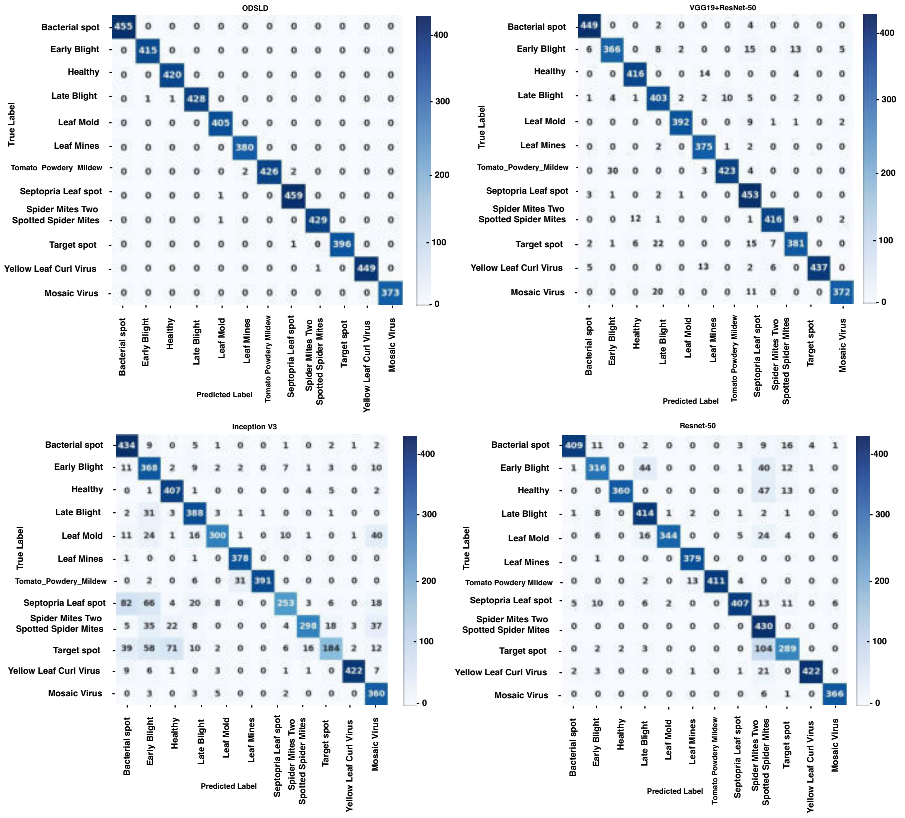


Fig. 4.5: Consequence matrices of ODSLD and existing DCNN model for tomato plant disease

The statistical performance metrics presented in Table. 4.5, show that the validation and testing loss score of the VGG19+ResNet-50 DCNN model is close to the ODSLD model compared to other DCNN models. The ODSLD and existing DCNN model validation and testing preciseness graph leveraging real-time images are portrayed in Fig. 4.4 (a) and (b) shows that the existing DCNN model learning accuracy is not improving because of the large unbalanced dataset, overfitting issue, and hyperparameter not optimised. Despite this, ODSLD model accuracy continues to improve with an increase in the number of testing and validation epochs. The ODSLD method performance validation is justified with the existing DCNN model based on the confusion matrix report depicted in Fig. 4.5.

Table 4.6: Comparative performance analysis of DCNN classification model for brinjal plantbrinjal

Method	Test Loss	Validation Loss	Learning rate	Optimizer
ODSLD	0.028	0.028	0.00080	SGD
ResNet-50	2.09	1.98	0.00045	Adam
Inception-V3	1.13	1.67	0.00073	Adam
VGG16	1.67	1.88	0.0034	Adam
CNN	2.79	2.38	0.0098	Adam
VGG19+ResNet-50	1.06	1.18	0.0076	Adam

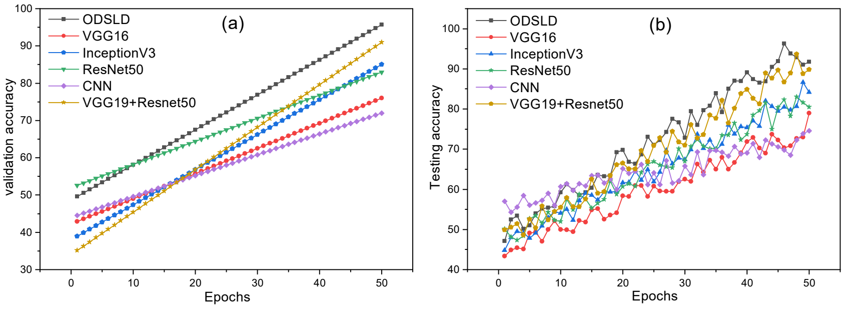


Fig. 4.6: Testing and validation accuracy and loss for brinjal plant using real time data

Inception V3, VGG19+ResNet-50 works better with large datasets [33], whereas proposed ODSLD model prediction is more accurate when tested and validated using large and small real-time image data. According to the confusion matrix shown in Fig. 4.5, ODSLD model able to recognise disease better than other existing DCNN models on real-time datasets. Despite this, ODSLD has the following advantages over other DCNN models. First of all, ODSLD produces a model with a comparatively lighter weight of 160 MB, the model can learn with customized large and small datasets than the other DCNN models while reducing the overfitting problem. Based on the loss-lessening feature of the ODSLD model, it is possible to classify the images more accurately with an average of 98.23% accuracy as shown in Fig. 4.4. The confusion matrices in Fig. 4.5 highlight that although VGG19+ResNet-50 and other DCNN model efficiently classify only 5-6 classes, ODSLD accurately classifies 12 classes, including novel illness classifications of Leaf Mines disease.

The ODSLD model for the brinjal plant has been established with analogous DCNN training and tuning parameters discussed above, and the ODSLD model performs exceedingly well compared to preceding DCNN classification models. Statistical evaluations of the ODSLD models for brinjal are presented in Table. 4.6.

The ODSLD model validation and testing loss are found to be better than the other DCNN models based on Table. 4.6 evaluation metrics. Fig. 4.6 depicts the ODSLD model validation and testing precisions and loss graph using real-time images, and Fig. 4.7 depict the prediction confusion matrix report of ODSLD and other DCNN models for the brinjal plant.

In Fig. 4.6 (a) and (b), the ODSLD model has the highest validation and testing accuracy of 95.67% and an average loss score of 0.028 shown in Table. 4.6 for the classification of six classes of brinjal plant diseases using real-time data. Despite the brinjal plant dataset's smaller size, the ODSLD model is more effective at classifying disease types than VGG19+ResNet-50 or other DCNN models. The confusion matrices in Fig. 4.7 emphasize that existing DCNN model classifies only 2-3 classes, and ODSLD precisely classifies all 6 types of disease classes, including novel illness classifications of Ladybugs caused diseases.

To recognize the disease in real-time, the ODSLD model has been dumped using the Keras API package. On the edge gateway, the ODSLD model is loaded. As soon as the OETA predicts a disease, the ODSLD model will recognize the disease type for both plant in real time and transmit the disease status to the cloud layer application through MQTT publish / subscribe protocol for monitoring so that farmers can take appropriate

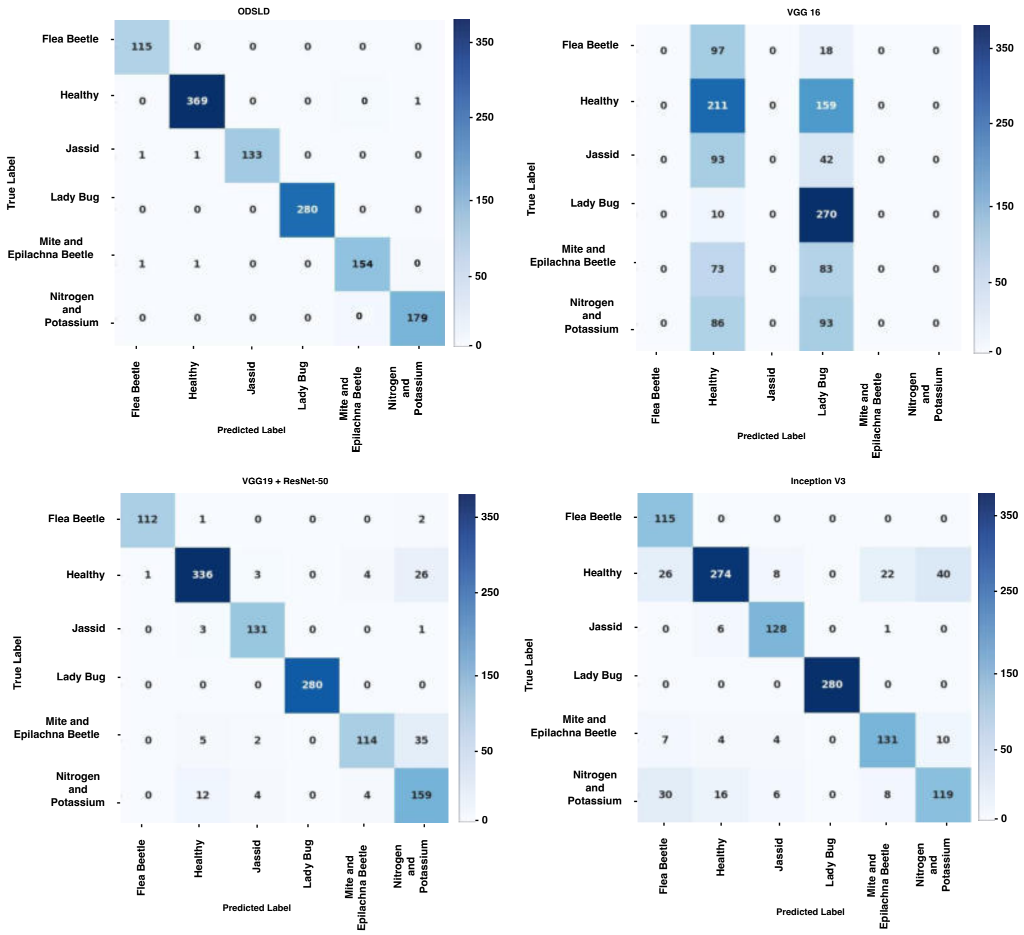


Fig. 4.7: Consequence matrices of ODSLD and existing DCNN model for brinjal plant disease

action as shown in the Fig. 4.8. Fig. 4.9 demonstrates the both plant disease detection monitoring graph. Date-wise disease detection status is displayed on the graph according to temperature, humidity, soil moisture, pH, nitrogen, phosphorus, and potassium.

Based on the aforementioned test outcomes and analysis, it seems to be assessed that the lightweight IoT edge-based hybrid (OETA + ODSLD) model is effective and efficient for detecting and diagnosing diseases for both leaves by overcoming the problem of existing smart leaves diseases method. Traditional pre-trained models, such as VGG16, ResNet50, and Inception v3, have many challenges in recognizing specific plant leaf diseases. Specifically, it is ineffective at capturing fine-grained data on both plant pathologies. To identify leaf disease, the DCNN model is hampered by its model size and inability to analyze comprehensive data. For example, variation in light, image quality, angles, resolution, or background noise. OETA is the best fit for disease prediction, considering data on soil nutrients and environmental conditions in Solanaceae plant fields to make an optimal decision. Proposed hybrid ODSLD method for both plant are the most suitable models for classifying disease type. The outcomes show that the most impoverished models for identifying illnesses in tomatoes are VGG16, ResNet50, and Inception V3 and for brinjal are inceptionV3, CNN and VGG16. In Table. 4.7, the overall proposed approach is compared with other existing approaches.

The comparison of the approaches shown in Table. 4.7 leads one to the finding that the suggested model for finding illnesses and categorization is automated, adaptable, and capable of determining the best decisions.

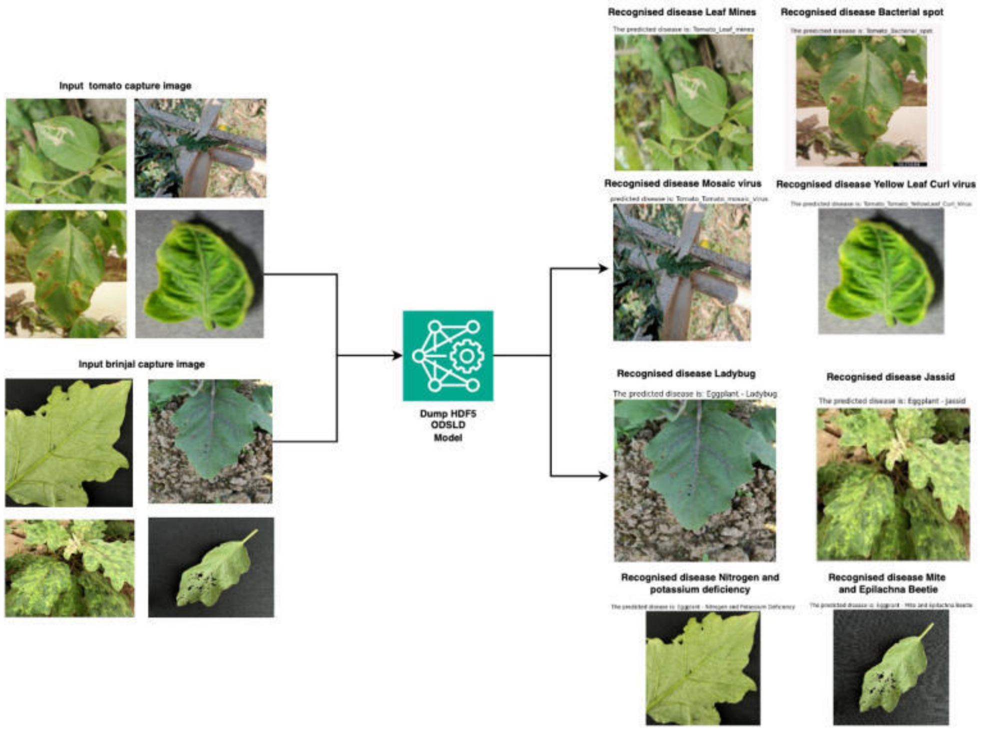

Input tomato capture image
Input brinjal capture image
Dump HDF5
ODSLD
Model
Recognised disease Leaf Mines
The predicted disease is: Tomato_Leaf_mines
Recognised disease Bacterial spot
The predicted disease is: Tomato_Bacterial_spot
Recognised disease Mosaic virus
predicted disease is: Tomato_Tomato_mosaic_virus
Recognised disease Yellow Leaf Curl virus
The predicted disease is: Tomato_Tomato_YellowLeaf_Curl_Virus
Recognised disease Ladybug
The predicted disease is: Eggplant - Ladybug
Recognised disease Jassid
The predicted disease is: Eggplant - Jassid
Recognised disease Nitrogen and potassium deficiency
The predicted disease is: Eggplant - Nitrogen and Potassium Deficiency
Recognised disease Mite and Epilachna Beetie
The predicted disease is: Eggplant - Mite and Epilachna Beetle

Fig. 4.8: Input captured image of detected disease with recognise output for Tomato and Brinjal plant

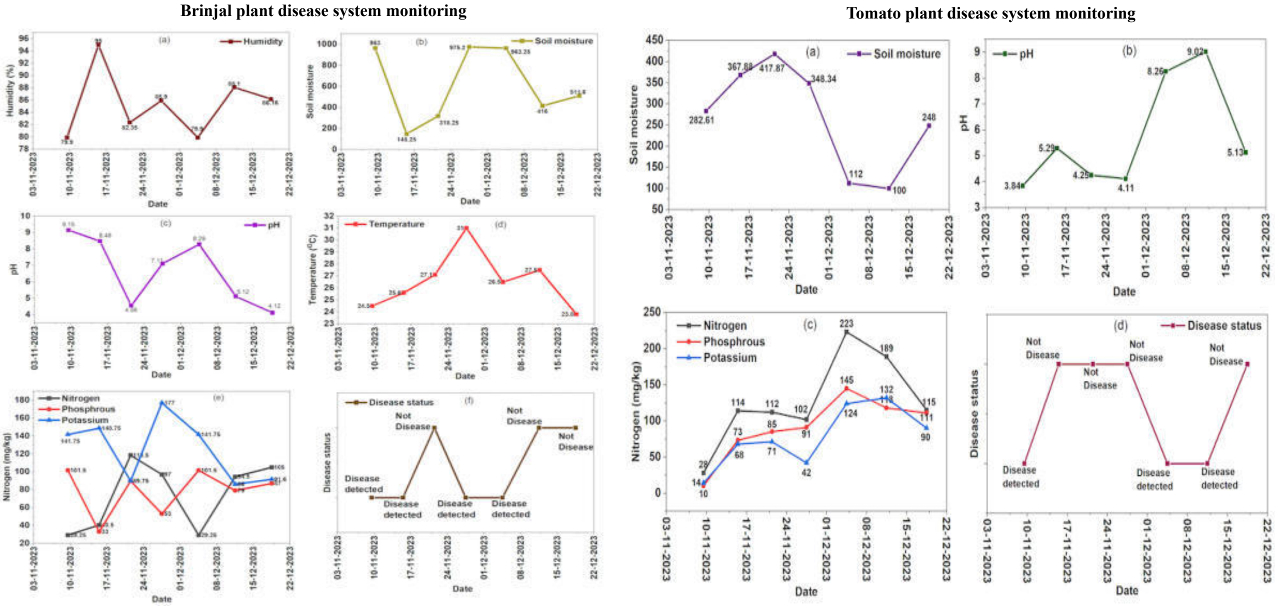


Fig. 4.9: Date wise disease detection status for both plant

Table 4.7: Comparative analysis summary

Reference	Method	Model performance	Customized system
[18]	SE inception DCNN based Disease recognition	Few number of disease are recognized	Large model size and absence of automated image capture, and disease identification system.
[20]	IoT ML	Disease detection depends on soil moisture content and climate only, and identifies fewer disease categories with lower predictive accuracy.	For automating the operation, the trained ML DCNN model is not dumped and absence of long range transmission.
[22]	IoT enabled with disease recognition	Since absence of real-time envision testing, the efficiency of the custom DCNN model is inadequate and non-adaptive.	Not deployed to evaluate model size and not fully automated and customized application.
Proposed framework	IoT Edge (OETA + ODSLD)	Detection of illness with 98.56% accuracy using various soil properties (SM, N,P, and K) and validation of disease recognition using real-time data, encompassing novel illnesses for both plant. Predict the disease each day and next days and wwek. Measure climate condition and soil health status.	To automate the system, the trained model is dumped with a size of 160 MB data for an brinjal and tomato plant. The system is able to communicate at long range wirelessly without any internet connectivity between the SLU to IGU.

5. Conclusion. This work presents models with implementation results for disease prediction, detection, and classification in tomato and brinjal plants. OETA framework is proposed for disease detection and prediction. ODSLD framework is proposed for classifying both plant leaf diseases. The LoRaWAN network transmits the tomato and brinjal field sensor data to an edge gateway. The OETA model normalizes the sensor data and detects and predicts disease. OETA's framework performs better for detecting diseases with an RMSE of 0.08 for tomatoes and 0.12 for brinjal plants. Whenever a disease, a real-time camera module captures the image of leaves and transmits it via LoRaWAN to the edge gateway. ODSLD pre-processes the captured image and classifies the disease. ODSLD model validation accuracy for tomato plants is 98.23% and for brinjal plants, it is 95.67%. The proposed method outperformed existing ML and DCNN models in a comparative analysis. This complete system will immensely assist small farmers in predicting, detecting, and classifying tomato and brinjal diseases. It will enable farmers to control diseases and alert them when to apply water, fertilizer, and pesticides to improve their harvest. In the future, other diseases of tomato and brinjal plant may be included in the proposed DCNN model to enhance model performance and provide assistance for automatically identifying disease remedies.

REFERENCES

[1] Bhattacharya, A., De, D, *AgriEdge: Edge Intelligent 5G Narrow Band Internet of Drone Things for Agriculture 4.0, In: Krause, P., Xhafa, F. (eds) IoT-based Intelligent Modelling for Environmental and Ecological Engineering. Lecture Notes on Data Engineering and Communications Technologies*, Springer, Cham, 67, 2021, doi: 10.1007/978-3-030-71172-6_3

[2] Mohanty, M.N., *Machine intelligence techniques for agricultural production*, *Smart Agriculture: Emerging Pedagogies of Deep Learning, Machine Learning and Internet of Things*, CRC press ,175, 2021, doi: 10.1201/b22627-13.

[3] Demilie, W.B, *Plant disease detection and classification techniques: a comparative study of the performances*, *J Big Data*, 11(5), 2024, doi: 10.1186/s40537-023-00863-9.

[4] W. L. Morris and M. A. Taylor,*The Solanaceous Vegetable Crops: Potato, Tomato, Pepper, and Eggplant*, *Encyclopedia of Applied Plant Sciences*, Elsevier, 55–58, 2017, doi: 10.1016/B978-0-12-394807-6.00129-5.

[5] Ghosh, Sunil Kumar, *Eggplant (Solanum melongena L.) and climate resilient agricultural practices*, *Climate Change Dimensions and Mitigation Strategies for Agricultural Sustainability (Volume II) edited by Suborna Roy Choudhury & Chandan Kumar Pand*, 2, 1-24, 2022.

[6] V. Sharma, A. K. Tripathi, and H. Mittal, *Technological Advancements in Automated Crop Pest and Disease Detection: A Review & Ongoing Research*, *2022 International Conference on Computing, Communication, Security and Intelligent*

Systems (IC3SIS), Kochi, India, IEEE, 1–6, 2022, doi: 10.1109/IC3SIS54991.2022.9885605.

[7] R. D. Devi, S. A. Nandhini, R. Hemalatha and S. Radha, *IoT Enabled Efficient Detection and Classification of Plant Diseases for Agricultural Applications*, *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India*, IEEE, 447-451, 2019, doi: 10.1109/WiSPNET45539.2019.9032727.

[8] Gupta, A., Nahar, P, *Classification and yield prediction in smart agriculture system using IoT*, *J Ambient Intell Human Comput*, 14, 10235–10244, 2023, doi: 10.1007/s12652-021-03685-w.

[9] B. Charbuty and A. Abdulazeez, *Classification Based on Decision Tree Algorithm for Machine Learning*, JASTT, 2(1),20 - 28, Mar. 2021, doi: 10.38094/jastt20165.

[10] Q. H. Cap, H. Uga, S. Kagiwada, and H. Iyatomi, *LeafGAN: An Effective Data Augmentation Method for Practical Plant Disease Diagnosis*, *IEEE Trans. Automat. Sci. Eng*, 19(2), 1258–1267, Apr. 2022, doi: 10.1109/TASE.2020.3041499.

[11] C. Wang, P. Du, H. Wu, J. Li, C. Zhao, and H. Zhu, *A cucumber leaf disease severity classification method based on the fusion of DeepLabV3+ and U-Net*, *Computers and Electronics in Agriculture*, 189, 106373, Oct. 2021, doi: 10.1016/j.compag.2021.106373.

[12] Rossy Nurhasanah, Lira Savina, Zul Mahadi Nata, and Ilham Zulkhair, *Design and Implementation of IoT based Automated Tomato Watering System Using ESP8266*, *Journal of Physics: Conference Series*, 1898, 012041, 2021, doi: 10.1088/1742-6596/1898/1/012041.

[13] Shamshiri, Redmond R., James W. Jones, Kelly R. Thorp, Desa Ahmad, Hasfalina Che Man and Sima Taheri, *Review of optimum temperature, humidity, and vapour pressure deficit for microclimate evaluation and control in greenhouse cultivation of tomato: a review*, *International Agrophysics*, 32, 287 - 302, 2018, doi: 10.1515/intag-2017-0005.

[14] Radicetti, Emanuele, Riccardo Massantini, Enio Campiglia, Roberto Mancinelli, S. Ferri, and Roberto Moscetti, *Yield and quality of eggplant (Solanum melongena L.) as affected by cover crop species and residue management*, *Scientia Horticulturae*, 204, 161-171, 161-171, 2016, doi: 10.1016/j.scienta.2016.04.005.

[15] Reichardt, K., Timm, L.C, *How Plants Absorb Nutrients from the Soil*, *In: Soil, Plant and Atmosphere*, Springer, Cham, 313–330, 2020, doi: 10.1007/978-3-030-19322-5_16.

[16] M. Kumar, A. Kumar and V. S. Palaparthy, *Soil Sensors-Based Prediction System for Plant Diseases Using Exploratory Data Analysis and Machine Learning*, *IEEE Sensors Journal*, 21(16), 17455-17468, 2021, doi: 10.1109/JSEN.2020.3046295.

[17] Haque MR, Sohel F, *Deep Network with Score Level Fusion and Inference-Based Transfer Learning to Recognize Leaf Blight and Fruit Rot Diseases of Eggplant*, *Agriculture*,12(8), 1160, 2022, doi: 10.3390/agriculture12081160.

[18] Li, Zhenbo, Yongbo Yang, Ye Li, RuoHao Guo, Jinqi Yang, and Jun Yue, *A solanaceae disease recognition model based on SE-Inception*, *Computers and Electronics in Agriculture*, 178, 2020, 105792, doi: 10.1016/j.compag.2020.105792.

[19] Jayanthi, M. G., Shashikumar, D. R, Preethi. S, *Eggplant leaf disease detection and segmentation using adaptively regularized multi Kernel-Based FuzzyC-Means and Optimal PNN classifier*, *Indian Journal of Computer Science and Engineering*, 13(5), 1542-1558, 2022, doi: 10.21817/indjcse/2022/v13i5/221305073.

[20] Saiqa Khan and Meera Narvekar, *Disorder detection of tomato plant(solanum lycopersicum) using IoT and machine learning*, *Journal of Physics: Conference Series*, IOP Publishing, 1432, 012086, 2020, doi:10.1088/1742-6596/1432/1/012086.

[21] Saad, Izazul Haque, Md Mazharul Islam, Isa Khan Himel, and Md Jueal Mia, *An automated approach for eggplant disease recognition using transfer learning*, *Bulletin of Electrical Engineering and Informatics*, 11(5), 2789-2798, 2022, doi: 10.11591/eei.v11i5.3575.

[22] Sowmiya, M., and S. Krishnaveni, *IoT enabled prediction of agriculture's plant disease using improvedπ quantum whale optimization DRDNN approach*, *Measurement: Sensors*, 27, 100812, 2023, doi: 10.1016/j.measen.2023.100812.

[23] Agarwal, Mohit, Abhishek Singh, Siddhartha Arjaria, Amit Sinha, and Suneet Gupta, *ToLeD: Tomato leaf disease detection using convolution neural network*, *Procedia Computer Science*, 167, 293-301, 2020, doi: 10.1016/j.procs.2020.03.225.

[24] T. Chen, D. Eager and D. Makaroff, *Efficient Image Transmission Using LoRa Technology In Agricultural Monitoring IoT Systems*, *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Atlanta, GA, USA*, pp. 937-944, 2019, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00166.

[25] Dong, M., Sha, H., Lu, M., Ma, Z., *ULIC: Ultra Lightweight Image Coder on Wearable Devices In: Zhai, G., Zhou, J., Ye, L., Yang, H., An, P., Yang, X. (eds) Digital Multimedia Communications. IFTC 2023. Communications in Computer and Information Science. Springer, Singapore*, 2067, 2024, doi: 10.1007/978-981-97-3626-3_8.

[26] Y. -H. Tsai and T. -C. Hsu, *An Effective Deep Neural Network in Edge Computing Enabled Internet of Things for Plant Diseases Monitoring,2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA*, 695-699, 2024, doi: 10.1109/WACVW60836.2024.00081.

[27] Nguyen T-H, Nguyen T-N, Ngo B-V. *A VGG-19 Model with Transfer Learning and Image Segmentation for Classification of Tomato Leaf Disease*, *AgriEngineering*, 4(4), 871-887, 2022. doi: 10.3390/agriengineering4040056.

[28] G. Huang, Z. Liu, G. Pleiss, L. v. d. Maaten and K. Q. Weinberger, *Convolutional Networks with Dense Connectivity*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 8704-8716, 2022, doi: 10.1109/TPAMI.2019.2918284.

[29] Motawani A, Khan. Q *Tomato Leaves Dataset*, *Kaggle*, Access on November 2023 https://www.kaggle.com/datasets/ashishmotwani/tomato?resource=download

[30] Dutta. G *Eggplant Diseases Detection*, *Kaggle*, Access on December 2023 https://www.kaggle.com/datasets/gauravduttakiit/eggplant-diseases-detection

[31] Loyani, Loyani, *Segmentation-based Quantification of Tuta absoluta's Damage on Tomato Plants*, *Smart Agricultural Technology*, 7, 100415, 2024, doi: 10.1016/j.atech.2024.100415.

[32] Shahinfar, Saleh, Paul Meek, and Greg Falzon, *How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring*, *Ecological Informatics*, 57, 2020, 101085, doi: 10.1016/j.ecoinf.2020.101085.

[33] Cao, J., Yan, M., Jia, Y. et al, *Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals*, *EURASIP J. Adv. Signal Process*, 49, 2021, doi: 10.1186/s13634-021-00740-8.

Edited by: Vadivel Ayyasamy
Special issue on: Internet of Things and Autonomous Unmanned Aerial Vehicle Technologies
for Smart Agriculture Research and Practice
Received: Apr 29, 2024
Accepted: Oct 8, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 114–122, DOI 10.12694/scpe.v26i1.3670

CYBERGUARD: A FORTIFIED MESSAGE AUTHENTICATION PROTOCOL WITH DIGITAL SIGNATURES IN NS-2 SIMULATION ENVIRONMENT

PRASHANT KUMAR[*], DIMPLE SETHI[†], GULSHAN SHRIVASTAVA[‡], AND NEERAJ SRIVASTAVA[§]

Abstract. Delving into the dynamic realm of communication networks, the Network Simulator-2 (NS-2) emerges as a versatile, object-oriented, and event-driven mirror, analysing the intricacies of computer network. NS-2, an open-source, is a strong tool for protocol innovation that supports numerous routing protocols designed for both wired and wireless networks in addition to the TCP/IP protocol suite. However, the security frontier has remained a latent challenge amidst its diverse protocol array. This research defiantly explores the security void in NS-2 and presents a novel solution by adding a state-of-the-art security module. Elevating NS-2's prowess, this module seamlessly integrates message integrity and sender authentication features, injecting a much-needed security boost. We meticulously detail the inner workings of the security modules, the innovative processes deployed, and the simulation and implementation intricacies within NS-2. The limelight is on the propagation of sender authentication protocols and the embodiment of message integrity in wired communication networks. Network Animator steps onto the stage to make the simulation journey more vivid, providing an engaging visual narrative. At its core, this module strives to democratize the integration of Digital Signature features, carving a path toward a more secure NS-2 landscape. This paper augments NS-2's resilience and charts new territories in the dynamic intersection of communication networks and cutting-edge security protocols.

Key words: Network Simulator-2, Decryption/Encryption, Integrity of Message, Sender Authentication, Asymmetric Key Cryptography, Network Layer Security, RSA Algorithm

1. Introduction. Network Simulator-2 is an event-driven open-source simulator developed primarily for research [15, 11, 8]. Experimenters can add additional modules and functions to NS-2 to meet their requirements. NS-2 comprises C++ and object-oriented Tool Command Language (OTCL) [7, 6]. C++ describes the inner workings (backend) of the simulation objects, whereas oTcl starts the simulation by building and configuring the objects and scheduling discrete events (frontend). NS-2 supports protocols from several levels, such as (the most recent version of NS-2 is version 2.35). Network Simulator-2 (NS-2) is a powerful, event-driven, open-source simulator expressly crafted for research endeavors. It offers a versatile platform wherein experimenters can seamlessly integrate new modules and functionalities to tailor the simulation to their specific requirements. NS-2 operates in two languages: C++ defines the internal mechanisms (backend) of simulation objects, while the object-oriented Tool Command Language (OTCL) orchestrates simulation setup by assembling and configuring objects, as well as scheduling discrete events (frontend). The latest version, NS-2 version 2.35, encompasses a comprehensive suite of standard protocols across various layers. From CSMA/CD at the data link layer to FTP, TELNET, DNS, and HTTP at the application layer, and encompassing both multicast and unicast routing protocols for wired networks, along with DSDV, DSR, and AODV for wireless networks, NS-2 has become a repository of advanced protocols embraced by the research community.

This paper introduces an innovative approach to enhance NS-2's capabilities by incorporating a security module. The proposed module facilitates the integration of message integrity and sender authentication functions. A novel packet format is constructed using a class derived from an NS-2 built-in class to instantiate a new protocol at the IP layer. This new class includes encryption/decryption functions for the data field in the data packet, ensuring sender authentication. The integrity of data during communication is confirmed using

[*] SCSET, Bennett University (The Times Group), Plot No 8-11, Techzone 2, Greater Noida, Uttar Pradesh 201310, India (prashant.mnnit10@gmail.com).

[†] SCSET, Bennett University (The Times Group), Plot No 8-11, Techzone 2, Greater Noida, Uttar Pradesh 201310, India (dimple9203@gmail.com)).

[‡] SCSET, Bennett University (The Times Group), Plot No 8-11, Techzone 2, Greater Noida, Uttar Pradesh 201310, India (Corresponding author, gulshanstv@gmail.com).

[§] Department of Mathematics, Agra college, Agra, Uttar Pradesh 282001, India (rssneeraj@gmail.com).

a hash function. The encryption/decryption and hash algorithms employed are RSA and MD5, respectively, utilizing TCL, AWK, and C++ as programming languages. The simulation environment necessitates NS-2 version 3.34, a GCC compiler, and a C++/C editor, all running on a personal computer with the Red Hat Enterprise Linux (RHEL) operating system. The subsequent sections of the paper delve into a comprehensive review of related work, the cloning of wired networks, encryption/decryption algorithms, fundamental exchange mechanisms, and the proposed scheme. The modifications are then registered, and the mirroring/simulation environment is delineated. The paper concludes with an evaluation of the approach's performance and security aspects, culminating in a comprehensive wrap-up of the findings.

2. Related Work. The TCP/IP protocol stack or different layers of the OSI protocol have different protocols implemented in NS2. For a wired network, application layer protocols include telnet, FTP, and HTTP; transport layer protocols include UDP, TCP, and SCTP; IP layer protocols include Ping, IP, and so on; and data link layer protocols such as CSMA/CD, which were created by scientists and eventually included in NS2. Similarly, multiple protocols at various OSI or TCP/IP protocol stack layers have been suggested for wireless networks, including TORA, AODV, DSDV, DSR [19] and DYMO [14].

Subsequently, NS-2 integrated these procedures. Recent research endeavors have concentrated on devising new protocols for both wired and wireless networks [12], with some focusing on enhancing or reforming existing protocols. However, there remains a need for more extensive efforts to amalgamate security measures with data transmission in NS2. Scholars like JIANG Hong, YU Qing-song, and LU Hui from East China Normal University in Shanghai, China, have directed their research efforts in this direction. Their methodologies contribute to bolstering Ethernet security by introducing a group-based MAC critical selection procedure (GKSP) tailored for large Ethernet networks. They have implemented security measures at the data link layer, wherein security concerns are addressed hop-by-hop. In our security module, we uphold security at the IP layer, as it aligns with various application requirements, thereby ensuring end-to-end security oversight. The security module we've introduced aims to address data integrity and sender authentication concerns at the network layer, thus enhancing and integrating data security functionalities within NS-2. Leveraging freely available resources and adopting the same programming platform as NS-2, based on RHEL [10], has guided our approach. Creating a novel security module for NS2 represents a significant stride toward addressing the need for inherent security capabilities within the simulation tool. This module introduces crucial encryption, decryption, and sharing capabilities vital for ensuring secure communication across diverse applications. This paper endeavors to augment NS2's simulation prowess in modeling secure communication scenarios by prioritizing reinforcing security measures. In another study [17], researchers assess the performance of four MANET routing protocols under various CBR and TCP traffic patterns using NS2. While AODV outperforms CBR traffic, OLSR exhibits superiority in handling TCP traffic. OLSR showcases optimal performance in managing multimedia/TCP traffic, making it well-suited for internet traffic scenarios. In [4] the authors introduce a communication system for microgrids (MGs) using NS2. Each MG includes a central controller and multiple distributed generation units with local controllers. The system facilitates data exchange between these controllers and sensors/actuators, employing low-cost ZigBee devices. The system integrates duplicate acknowledgment and data management schemes to optimize data transfer despite ZigBee's limitations. It also ensures intelligent data routing in case of path or device failures. Performance metrics like Packet Delivery Ratio (PDR), throughput, and end-to-end communication time are evaluated to validate system effectiveness. The routing protocol selection in wireless networks, specifically on Mobile Adhoc Networks (MANETs) where throughput and minimal delay are critical [2]. Their study evaluates the performance of Adhoc On-demand Distance Vector (AODV) and Destination Sequenced Distance Vector (DSDV) protocols using an NS2 simulator, emphasizing parameters like packet throughput, jitter, and end-to-end delay. By designing a wireless network of mobile nodes with defined parameters, the researchers aim to provide a detailed comparative analysis to aid in selecting the most suitable protocol for optimal network performance.

The necessity for energy-efficient protocols in Wireless Sensor Networks (WSNs), focusing on the "Sensor-Medium Access Control" (S-MAC) protocol for its ability to reduce sensor node energy consumption is discussed in [9]. It highlights the challenge of preserving energy while meeting application demands in battery-constrained sensor nodes. The study introduces a novel energy-efficient clustering algorithm based on the LEACH protocol, comparing it with other clustering protocols such as LEACH-C, MTE, and Stats-Clustering through NS2 simulations. Additionally, the paper analyzes S-MAC using Network Simulator NS2.

NS2, developed by UC Berkeley, is an open-source simulator for Internet Protocol Networks. While it produces ASCII-formatted trace files capturing simulation events, analyzing these files can be challenging due to their textual format and large size. Q-Analyze, a trace file analysis tool, simplifies data extraction for performance metrics measurement and generates user-friendly simulation reports. Operating through three layers, Q-Analyze streamlines network performance study by providing intuitive GUI and focusing on algorithm development rather than data processing and metric calculation. NS2, developed by UC Berkeley, is an open-source simulator for Internet Protocol Networks. While it produces ASCII-formatted trace files capturing simulation events, analyzing these files can be challenging due to their sizeable textual format. Q-Analyze, a trace file analysis tool, simplifies data extraction for performance metrics measurement and generates user-friendly simulation reports. Operating through three layers, Q-Analyze streamlines network performance study by providing intuitive GUI and focusing on algorithm development rather than data processing and metric calculation [3].

3. Background.

3.1. Simulation of Wired Network. In NS-2, our objectives are achieved in two ways:

- The Otcl script has to be changed, but the built-in network modules in NS-2 do not need to be changed when they are sufficient to achieve the mirroring/simulation goal [11, 16, 20].
- If the pre-installed network modules/components are insufficient for mirroring/simulation, a new module/component must be made, or the current modules must be modified [21, 1, 13]. In other words, NS2 should be expanded by introducing a new Otcl class.

Adjustments must be made to the Otcl script to execute the simulation. Utilizing the existing components suffices for simulating the basic Mobile IPv4 protocol. However, for tailored functionalities such as specialized applications, business flows, agents, linkages, routing, and node models, thorough verification and compilation of these components are imperative to ensure seamless integration [5]. Post-simulation, analyzing trace files yields invaluable insights. Additionally, monitoring the network simulation process can be facilitated by utilizing NAM. The insights gleaned from simulation analysis are instrumental in discerning whether tweaks to the configuration topology and business simulation are warranted to initiate further simulations to achieve desired simulation outcomes.

3.2. Digest Generation using MD5 Algorithm. Professor Ronald L. Rivest from MIT is credited with creating MD5 [18]. MD5, an algorithm designed to generate a 128-bit fingerprint or message digest for any message, regardless of its length, is widely recognized for its computational in-feasibility in producing two messages with the same message digest or any message with a predetermined goal message digest. MD5 securely compresses large files in digital signature applications before encrypting them with a private key using a public-key cryptosystem such as RSA. Renowned for its reliability, MD5 surpasses checksum and many other widely used techniques in verifying data integrity. The core MD5 algorithm operates on a 128-bit state, segmented into four 32-bit words initialized to specific fixed constants. It iteratively processes each 512-bit message block, modifying the state. This processing occurs in four rounds, each comprising 16 similar operations: modular addition, left rotation, and a non-linear function known as F.

3.3. RSA Algorithm for Data Security. The RSA algorithm [18] was created by Ron Rivest, Adi Shamir, and Len Adleman, who created it in 1977. The RSA cryptosystem is the world's most used public key cryptography algorithm. It allows communication to be encrypted without exchanging a secret key individually. The RSA technique may apply to public key encryption and digital signatures. Its security relies on the difficulty of factoring huge numbers. Party A can deliver encrypted communication to Party B without exchanging secret keys. A encrypts the message with B's public key, and B decrypts it with only the private key he knows. RSA can also be used to sign a message, so A can sign a message using their private key, and B can verify it using A's public key.

Sender Side Encryption Process: Sender A adheres to these guidelines:

1. obtains the public key of recipient B i.e. (m, f).
2. a positive integer that symbolises the textual message pt, with $1 < pt < m$.

3. uses the formula to determine the cipher text d $d = pt^f \mod m$.
4. sends the encrypted text to B.

Receiver Side Decryption Process: Recipient B adheres to these guidelines:

1. uses the cipher text d and their private key (m, e) to calculate pt, using the formula $pt = d^e \mod m$.
2. extracts the plaintext from the representation of an integer (pt).

4. Proposed Work. To ensure message integrity and sender authentication, the self-defined security protocol Class, or *packet_sec*, implements hash and decrypt/encrypt functions. Furthermore, the security protocol keeps a sequence number seq for each node. If the sender delivers a packet, seq is increased by one and adds to the packet header information, allowing the receiving node to reorganize the packets. The general protocol believes a source file, *packet_sec.h/.cc* (the protocol's solution and realization) should be created. Algorithm 1 illustrates the definition of a security protocol in which the command and recv functions are inherited from the Agent class. The inter-layer communication mentioned is detailed within the Tcl code of NS2. The class definition referred to is available in *packetsec.h.* To enable this security protocol in Tcl, NS must recognize it as an Agent. The standard procedure to meet these criteria involves defining the packet sec class in C++ within Agent/packet sec. This facilitates the modification of the Tcl code and the definition of the security protocol. An inheritance relationship exists between Agent and packet sec, with packet sec inheriting from Agent.

Algorithm 1 Class `proto_sec_anal`: Public Agent

```
Class proto_sec_anal: Public Agent {
Public:
  void accept(Packet * k, Handler *)
  $p$: Packet accepted by the receiver
  Handler for processing
  int command(int argc, const char * const * argv)
  argc: Number of command-line arguments
  argv: Array of command-line argument strings
Private:
  uint32t serial_no
}
```

4.1. Implementation of the Digital Signature and Hash algorithms . Sender A computes the hash and acquires a digital signature as follows: In Algorithm 2, generating a hash on data retrieved from the TCL file and obtaining a digital signature for sender authentication is depicted. Initially, a packet is created using the authorized function allocpkt(). Subsequently, the packet header is accessed by creating an object named hdr with the type hdrpacketsec. The sending time is also provided in the send_time variable of the packet header, which functions as a timestamp. To obtain the hash value, the data, and its length are fed into the hashing algorithm, resulting in the hash value. This value is then utilized to verify the integrity of the message. Afterward, the RSA algorithm is applied to the data and passed to the decryption/encryption function, resulting in encrypted data. This encrypted data is assigned to the data variable in the packet header. Finally, these values are stored in packet header variables, and the packet is transmitted to the receiver [10].

Packet Verification by the Receiver. The process depicted in Algorithm 3 ensures that receiver B initiates a new hash generation and verifies the signature. Upon receiving the packet, the receiver initially stores the transmission time, sequence number, hash value, and encrypted contents in fresh variables. Subsequently, the receiver decrypts the data using the RSA method, forwarding it to a decryption function. If necessary, the function returns the original data. The decrypted data is then hashed using a hashing algorithm. Upon conclusion, the received hash value is compared with the newly computed hash value. If both values match, a signed acknowledgment is returned; otherwise, the message is adjusted.

The receiver sends an ack to relay the result. The approach presented in Algorithm 4 demonstrates how the receiver conveys the signature verification result. Initially, the receiver generates a new packet using the allocpkt() method and accesses the packet's header via the hdrret object. Later, ret variable is set to 1. So

Algorithm 2 Message Digest and Digital Signature Generation

```
if (strcmp(argv[1], "transmit") == 0) then
    // Make a one new packet
    Packet * packet = allocpkt()
    // Incorporate security packet header in the newly created packet
    packethdrsec * packet_hdr = packethdrsec::access(packet)
    // To let the receiving node know that it must send and acknowledge packets, set the ’bak’ variable to 0.
    packet_hdr->bak = 0
    packet_hdr->sno = sno++
    // Put the time now in the ”transmit_time” field.
    packet_hdr->transmit_time = Scheduler::instance().clock()
    // Move the copied info to the header.
    strcpy(packet_hdr->info, argv[2])
    //@@@@@@@@@@@@@ Hashing Operation @@@@@@@@@@@@@//
    packet_hdr->msgdigest = hashing (packet_hdr->info, (unsigned int) strlen (packet_hdr->info))
    //@@@@@@@@@@@@@ Securing the data @@@@@@@@@@@@@@//
    data_encryption(packet_hdr->info)
    printf("Digital Signature Generated:")
    //@@@@@@@@@@@@@ Send a packet @@@@@@@@@@@@@@@@//
    transmit(packet, 0)
    return (TCL OK)
end if
```

the receiver would not send another echo. Thereafter, report the transmitting time to the packet header’s send_time variable. Finally, result is saved in the packet header’s data field.

4.2. TCL File Modifications. TCL serves as a scripting language employed in developing network units and components. For our simulation, six nodes are established, four of which have security agent capabilities and are connected. Their connections are in the following order: node_0 to node_5 and node_1 to node_4. Later, invoke the function that sends the data. Lastly, delve into the recv function, which retrieves the variable’s value and exhibits it on the Linux terminal upon successful execution. The Tcl file about digital signature verification and hash creation is depicted in Algorithm 5.

5. Security Module Registration. Marc Greis’ example [18] illustrates introducing a new protocol to NS2. Initially, a new packet Class is created in the application folder ../apps. Following this, the packet name is inserted into the packet.h header file of NS2. Subsequently, modifications are made to the makefile to accommodate the creation of the new Class. Each newly created packet must be defined at the TCL layer by adding the default packet size value and name to the ns-default.tcl file. Finally, an entry for the newly created packet is added to the ns-packet.tcl file. Recompiling NS-2 will then prepare the new packet for simulation. A new packet type is devised for transmitting data.

6. Security and Permormance Analysis. After testing, analyze the security protocol’s performance using experimental data statistics. Figure 6.1 shows the packet transmission status using the security protocol [19]. It may be seen in Figure 6.1. In security protocols, data is sent via unicast (Algorithm 6). When a node delivers a data packet, it unicasts it to directly linked nodes. If a node accepts a packet but has not yet transmitted it and the destination is not itself, it will forward it in unicast format. The procedure will continue until the data packet reaches its destination.

6.1. Environment Simulation. In the environment illustrated in Figure 6.1, NAM constructs a network with six nodes interconnected by a 5 Mbps full duplex link. Security agents are deployed on nodes n1, n4, and n5. Communication transpires between nodes n0 and n5 and nodes n1 and n4 via the link between n2 and n3. A drop-tail queue of 100 is employed between nodes n2 and n3. Following one minute of operation, a trace file is generated. Subsequently, the trace file undergoes processing using a text processing language like awk to derive the desired output.

Algorithm 3 Signature verification at destination

```
if packet_hdr->bak == 0 then
  double ttime = packet_hdr->transmit_time
  //@@@@@@@@@@ Packet data encryption/Decryption @@@@@@@@@@//
  char actual_info[128]
  char secure_info[128]
  strcpy(secure_info, packet_hdr->info)
  // Make a copy of the original packet's data
  strcpy(actual_info, packet_hdr->info)
  int accept_seq = packet_hdr->sno
  //@@@@@@@@@ Present the packet to the recipient node @@@@@@@@@//
  char var[105]
  unsigned int newdigest
  char verified_outcome[50]
  decryption(actual_info)
  newdigest = hashing(actual_info, strlen(actual_info))
  if newdigest == packet_hdr->msgdigest then
    printf("Signature matched")
    strcpy(verified_outcome, "Sender Authenticated")
  else
    printf("Sender did not sign the msg %dnn", newdigest)
    strcpy(verified_outcome, "Message Altered")
  end if
  // Drop the packet
  Packet::free(packet)
end if
```

Algorithm 4 Receiver sends ACK packet to confirm delivery of message

```
// Make a one new Packet
Packet * packetack = allocpkt()
// Access the new packet's header.
packethdrsec * packet_hdrack = packethdrsec::access(packetack)
packet_hdrack->bak = 1
// Enter the accurate value in the transmit time field
packet_hdrack->transmit_time = ttime
packet_hdrack->receive_time = Scheduler::instance().clock()
packet_hdrack->sno = accept_seq
strcpy(packet_hdrack->info, verified_outcome)
send(packet_hdrack, 0)
```

6.2. Analysis of Security Protocol. Figure 6.1 depicts the transmission of a unicast data packet. After executing the security procedure, Algorithm 6 illustrates concise results. Initially, node n0 transmits the data to node n5 utilizing the hash and RSA technique. Node n5 then receives the packet and decrypts the message using the RSA technique, retrieving the original contents. Subsequently, a hash of the decrypted data is generated and compared to the received hash value. If a match is found, the node acknowledges receipt of a signature-validated message; otherwise, a message indicating data alteration is transmitted. Similarly, node n1 communicates with node n4. This process iterates in reverse order accordingly. The time required for encryption, decoding, and hash calculation was also computed.

7. Conclusion. In summary, this study delves into the simulation intricacies of a wired network, incorporating two pivotal protocols: message integrity and sender authentication. We meticulously expound upon the simulation methodology within NS2, elucidating our analytical approach to interpreting the results. The visualization of the simulation process is facilitated through NAM, while Awk emerges as the tool of choice for

Algorithm 5 Receiver sends ACK packet to confirm delivery of message

```
set q1 [new Agent/pkt_sec]
$ns attach-agent $m1 $q1
$q1 set class 2
set q2[new Agent/pkt_sec]
$ns attach_agent $m2 $q2
$q2 set class 2
set q3[new Agent/pkt_sec]
$ns attach-agent $m5 $q3
$q3 set class 3
set q4[new Agent/pkt_sec]
$ns attach-agent $m6 $q4
$q4 set class 3
// Link the two agents together
$ns connect $p0 $p3
$ns connect $p1 $p2
// Plan your events
for {set j 1}{$j<2} {incr j}{
set outcome [expr $j/2]
$ns at [expr $outcome + 0.04] "$q1 transmit thisispk"
$ns at [expr $outcome + 0.40] "$q2 transmit thisispks"
$ns at [expr $outcome + 0.80] "$q3 transmit thisis"
$ns at [expr $outcome + 1.20] "$q4 transmit thisis..."}
// function 'accept1' for the 'Agent/pkt_sec' class
Agent/pkt_sec instproc accept1 (from rtt) {
$self instvar node
puts "node [$node id] obtained a secret Key from
$from with trip-time $rtt ms"}
// function 'accept' for the 'Agent/Packet_sec' class
Agent/pkt_sec instproc accept (from rtt
mess originmess hash) {
$self instvar node
puts "node[$node id]collected_packet from
$from with trip-time $rtt ms - contend: $mess -
decrypted $originmess -hash: $hash"}
```

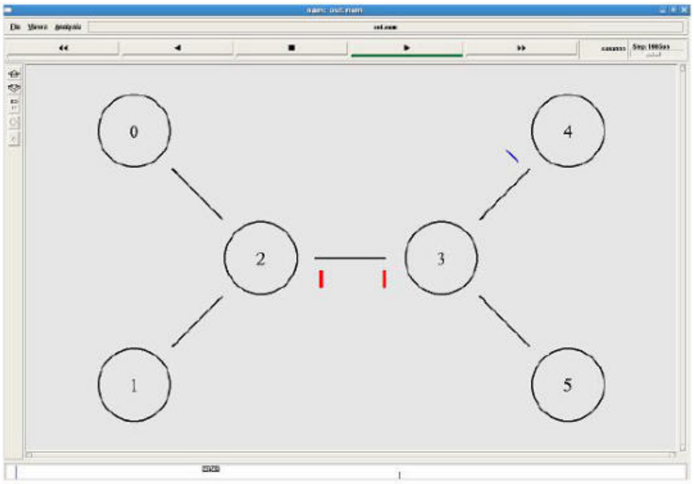


Fig. 6.1: Creating an environment using Network Animator Tool (NAM)

Algorithm 6 Outcome of security protocol after execution

```
Message transmit: "Thisispk" along with message digest using MD5 algorithm
Message Digest: 906d9a4206f569725ffb4b2b178baba6 & SIGNATURE CREATED
Message Digest Collected by Receiver: 906d9a4206f569725ffb4b2b178baba6
SIGNATURE IS VERIFIED
Node 5 obtained a packet from 0 with a 31.1 ms travel time. -
contend encrypted_data: sCcbIEYv8pe/9+XNCpFrTMW0Y -
decrypted: Thisispk
node 5 sent an ack packet to node 0 with a travel time of 62.2 ms. -
contend: SENDER_AUTHENTICATED
Message transmit "Thispks" along with message digest using MD5 algorithm
Message Digest: 453558b732457b31e8eb083afdf5435f and DIGITAL SIGNATURE CREATED
MD5 Digest at Receiver Side: 453558b732457b31e8eb083afdf5435f
SIGNATURE VERIFIED
With a journey time of 31.1 ms, node 4 received a packet from 1. -
contend encrypted_data: pMECBU04/RYvM94XwLrpzlsLPj1J6ORNxtsu1ExN
decrypted: Thisispks
Node 1 obtained a packet from 4 with a 62.2 ms travel time. -
contend: SENDER_AUTHENTICATED
```

trace file post-processing and nuanced comparisons. Expanding our investigation, we systematically manipulate network topologies and data flows in diverse configurations, consistently yielding closely aligned results. The successful integration of a security module into NS2 marks a significant milestone, enabling its versatile application across a spectrum of data security-intensive scenarios. This juncture affords us the capability to scrutinize and assess various applications post-implementation, delving into facets such as security overheads, packet loss, required bandwidth, throughput, and other pertinent metrics. Within the implemented module, the MD5 algorithm serves to generate a 128-bit digest, while the RSA algorithm assumes the role of our encryption mechanism. It is imperative to note that the flexibility exists to substitute the RSA algorithm with any public key algorithm; the current preference for RSA is rooted in its pragmatic simplicity. Emphasizing its specificity to wired networks, our future endeavors entail extending this module's support to encompass wireless networks.

Anticipating the future, our research trajectory extends beyond the current scope, entering a phase of advanced exploration and innovation in the realm of network security simulations. With a particular emphasis on the dynamic difficulties and developing threats in the cybersecurity landscape, we are dedicated to an unyielding pursuit of thorough studies that delve deeper into the complex details of applications post-implementation. Our commitment to excellence is reflected in the ongoing refinement and enhancement of our simulation methodology and analytical frameworks. This involves continuous integration of cutting-edge technologies, adaptive algorithms, and advanced cryptographic techniques to fortify the security module. Through a continuous feedback loop of simulation results and empirical insights, our aim is to fine-tune our models, ensuring they remain resilient in the face of emerging security threats. Furthermore, our future endeavors include the exploration of novel encryption algorithms and cryptographic primitives, seeking to elevate the security standards within network simulations. We anticipate delving into the realm of quantum-resistant cryptography, considering the evolving landscape of quantum computing and its potential implications for network security. In our quest for an evolved security paradigm, we are committed to extending the application of our module to diverse network architectures, including wireless networks. This expansion will involve adapting the existing security framework to the unique challenges posed by wireless communication, such as channel vulnerabilities and mobility issues. As proceed, collaborative efforts with industry partners, academic institutions, and cybersecurity experts will play a pivotal role. This collaborative approach will not only validate the robustness of our simulations but also foster a knowledge-sharing ecosystem, contributing to the collective advancement of network security research.

In conclusion, our research trajectory is characterized by a forward-looking perspective, driven by an unyielding commitment to advancing the field of network security simulations. Through continuous refinement, exploration of emerging technologies, and collaborative partnerships, we aspire to shape a future where network simulations serve as a cornerstone for developing resilient and adaptive cybersecurity solutions.

REFERENCES

[1] D. Adami, C. Callegari, D. Ceccarelli, S. Giordano, and M. Pagano, *Design, development and validation of an ns2 module for dynamic lsp rerouting*, in Computer-Aided Modeling, Analysis and Design of Communication Links and Networks, 2006 11th International Workshop on, 2006, pp. 72–77.

[2] N. Aggarwal, T. S. Chohan, K. Singh, R. Vohra, and S. Bahel, *Relative analysis of aodv & dsdv routing protocols for manet based on ns2*, in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, 2016, pp. 3500–3503.

[3] Q. Al-Shidi, A. Alburaiki, H. Shaker, and B. Kumar, *Q-analyze tool to detect malicious and black hole nodes in ns2 simulation for aodv*, in 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), IEEE, 2018, pp. 140–146.

[4] P. O. Bhalerao, *Communication system design and simulation for future micro grids in ns2*, in 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), IEEE, 2016, pp. 688–690.

[5] N. Glance, D. Snowdon, and J.-L. Meunier, *Pollen: using people as a communication medium*, Computer Networks, 35 (2001), pp. 429–442.

[6] M. Greis, *Tutorial for the network simulator 'ns'*. `http://www.isi.edu/nsnam/ns/tutorial/`, 2006.

[7] Z. Ishak, N. Din, and M. Jamaludin, *Ipqit: An internet simulation kit based on ns2*, in Telecommunications and Malaysia International Conference on Communications, 2007. ICT-MICC 2007., 2007, pp. 489–493.

[8] H. Jiang, Q.-s. Yu, and H. Lu, *Simulation and analysis of mac security based on ns2*, in Multimedia Information Networking and Security, 2009. MINES '09., vol. 2, 2009, pp. 502–505.

[9] K. H. Krishna, T. Kumar, and Y. S. Babu, *Energy effectiveness practices in wsn over simulation and analysis of s-mac and leach using the network simulator ns2*, in 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2017, pp. 914–920.

[10] P. Kumar, S. Maheshwari, and H. K. Dubey, *Development and validation on ns2 protocol for data security at network layer*, in International Conference on Computing, Communication & Automation, 2015, pp. 529–534.

[11] S. Kumar, R. Rathy, and D. Pandey, *Traffic pattern based performance comparison of two reactive routing protocols for ad hoc networks using ns2*, in Computer Science and Information Technology, 2009. ICCSIT 2009., 2009, pp. 369–373.

[12] G. Li and J. Chen, *The research of routing algorithms based on ns2 in mobile ad hoc networks*, in Software Engineering and Service Science (ICSESS), 2011., 2011, pp. 826–829.

[13] M. Qadeer, V. Sharma, A. Agarwal, and S. Husain, *Differentiated services with multiple random early detection algorithm using ns2 simulator*, in Computer Science and Information Technology, 2009. ICCSIT 2009., 2009, pp. 144–148.

[14] M. Q.-x. Qu and L. Xu, *Dymo routing protocol research and simulation based on ns2*, in Computer Application and System Modeling (ICCASM), 2010., vol. 14, 2010, pp. V14–41 – V14–44.

[15] Z. Qun and J. Wang, *Application of ns2 in education of computer networks*, in Advanced Computer Theory and Engineering, 2008. ICACTE '08., 2008, pp. 368–372.

[16] C. P. Reddy and C. P. Reddy, *Performance analysis of adhoc network routing protocols*, in Ad Hoc and Ubiquitous Computing, 2006. ISAUHC '06., 2006, pp. 186–188.

[17] S. J. Soni and J. S. Shah, *Evaluating performance of olsr routing protocol for multimedia traffic in manet using ns2*, in 2015 Fifth International Conference on Communication Systems and Network Technologies, IEEE, 2015, pp. 225–229.

[18] W. Stallings, *Cryptography and Network Security*, fourth edition ed., 2006.

[19] A. Tuteja, R. Gujral, and S. Thalia, *Comparative performance analysis of dsdv, aodv and dsr routing protocols in manet using ns2*, in Advances in Computer Engineering (ACE), 2010., 2010, pp. 330–333.

[20] S. Xu and Y. Yang, *Protocols simulation and performance analysis in wireless network based on ns2*, in Multimedia Technology (ICMT), 2011., 2011, pp. 638–641.

[21] S. Zhao, P. Wang, and J. He, *Simulation analysis of congestion control in wsn based on aqm*, in Mechatronic Science, Electric Engineering and Computer (MEC), 2011., 2011, pp. 197–200.

Edited by: Kavita Sharma
Special issue on: Recent Advance Secure Solutions for Network in Scalable Computing
Received: Apr 6, 2024
Accepted: Jun 12, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 123–135, DOI 10.12694/scpe.v26i1.3128

RESEARCH ON VEHICLE ROUTING PROBLEM WITH TIME WINDOW BASED ON IMPROVED GENETIC ALGORITHM

XU LI[*], ZHENGYAN LIU[†], AND YAN ZHANG[‡]

Abstract. This article conducts a detailed study on the vehicle routing problem with time window constraints. We constructed an objective function for the vehicle routing problem with time windows, established a mathematical model, and proposed an improved genetic algorithm to solve the problem. The algorithm first constructs a chromosome encoding method, designs a heuristic initialization algorithm to generate a better initial population, and determines the fitness function. During the operation of the algorithm, selection, crossover, and mutation operations are designed to generate offspring populations, enhancing the diversity of the population and avoiding premature convergence of the algorithm. Meanwhile, in order to improve the optimization and local search capabilities of genetic algorithms, this paper constructs a local search operation. Finally, the algorithm implements an elite retention strategy on the parent population and reconstructs a new population. We conducted simulation experiments on the algorithm using MATLAB and selected examples from the Solomon dataset for testing. The simulation experiment results have verified that the improved genetic algorithm is feasible and effective in solving vehicle routing problems with time windows.

Key words: vehicle routing problem, time window, genetic algorithm, local search

1. Introduction.

With the rapid development of the logistics industry, optimizing vehicle delivery routes has become increasingly urgent. How to improve logistics distribution efficiency and reduce distribution costs while meeting customer node needs has become an urgent problem that logistics distribution enterprises need to solve [8]. Vehicle routing problem with time windows (VRPTW) has become a hot research topic in recent years due to its strong practical significance [1, 12, 16, 2, 17]. The vehicle routing problem with a time window is an NP hard problem. If the problem is solved using an exact algorithm, it will take too long, while the solution quality obtained by traditional heuristic algorithms is not high. The swarm intelligence optimization algorithm is increasingly being applied to the solution of this problem due to its characteristics of parallelism, universality, and stability. Mst. Anjuman Ara et al. [3] proposed a hybrid genetic algorithm which incorporates three different heuristics for generating initial solution including sweep algorithm, time oriented heuristics and swap heuristic. Khoo Thau Soon et al. [9] proposed the parallelization of a two-phase distributed hybrid ruin-and-recreate genetic algorithm for solving multi-objective vehicle routing problems with time windows. Hongguang Wu et al. [18] proposed a hybrid ant colony algorithm based on ant colony algorithm and mutation operation. Guo Ning et al. [7] proposed an adaptive variable neighborhood search ant colony algorithm to solve the vehicle routing problem with soft time windows. Chen Ying et al. [4] proposed a hybrid particle swarm optimization algorithm based on hierarchical learning and differential evolution.

Due to the strong global optimization ability and good robustness of genetic algorithms, this paper proposes an improved genetic algorithm for solving vehicle routing problems with time windows. We first analyze in detail the characteristics of the vehicle routing problem with time windows and establish a problem model. In algorithm design, we construct a chromosome encoding method, design a heuristic initialization algorithm, and design selection, crossover, and mutation operations. Meanwhile, a local search operation is designed to address the drawback of genetic algorithms being prone to falling into local optima. Finally, simulation experiments are conducted to verify the effectiveness, reliability, and universality of the constructed model and the designed algorithm.

[*] School of Computer and Information Engineering, Fuyang Normal University, Fuyang, China

[†] School of Computer and Information Engineering, Fuyang Normal University, Fuyang, China (Corresponding author, zhyliu@fynu.edu.cn).

[‡] School of Computer and Information Engineering, Fuyang Normal University, Fuyang, China

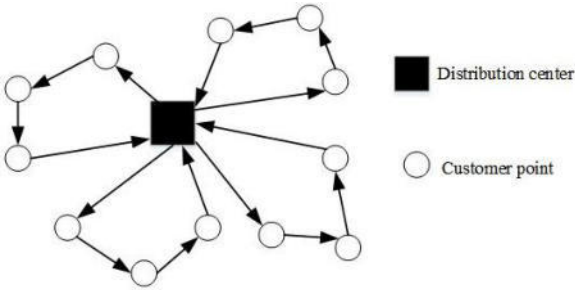


Fig. 2.1: Schematic diagram of vehicle routing problem

2. Vehicle Routing Problem (VRP).

2.1. Problem description. The distribution center distributes goods to customer nodes, and the number, location, and demand for goods at each customer node are known. Transport vehicles start from the distribution center, conduct orderly access to designated customer nodes, and meet some constraints, such as the demand for goods from customer nodes, the time to receive goods, vehicle load capacity limitations, etc., and finally return to the logistics distribution center. Reasonable distribution routes need to be designed to minimize vehicle routes, time consumption, and transportation costs [5, 14]. The schematic diagram of the vehicle routing problem is shown in Fig. 2.1.

2.2. Vehicle routing problem with time windows (VRPTW). The vehicle routing problem with time windows is based on the vehicle routing problem model, which adds a time window as a constraint condition to the customer node. VRPTW is an important branch of VRP, generally described as: several delivery vehicles depart from the distribution center, deliver goods to customer nodes in sequence, complete the delivery task, and return to the distribution center. During the delivery process, there can only be one vehicle serving the customer node, and it is required to deliver goods to the customer node within the specified time period [13]. If the vehicle arrives early or late, corresponding penalties will be imposed.

The vehicle routing problem with time windows can be further divided into soft time window vehicle routing problem and hard time window vehicle routing problem. The soft time window vehicle routing problem refers to the customer's request for logistics delivery vehicles to complete delivery tasks as soon as possible within the specified time period, otherwise they will be subject to corresponding penalty fees. The hard time window vehicle routing problem refers to the logistics delivery vehicle needing to complete the delivery task to the customer within the specified time period, otherwise the customer refuses to receive the goods.

3. Establishment of a Model for VRPTW.

3.1. Model assumptions. In order to facilitate algorithm design and problem research, the following reasonable assumptions are made before modeling [10].

1. The location of the distribution center and customer nodes is determined.
2. Delivery vehicles are of the same type, and each vehicle has the same load capacity.
3. The distance from the distribution center to customer nodes and the distance between customer nodes are known.
4. Assuming that the delivery vehicle departs from the distribution center at a time of 0.
5. Each delivery vehicle can serve multiple customer nodes, but each customer node has only one delivery vehicle serving it.
6. The demand for goods at each customer node is clear.
7. The customer node's cargo demand is less than the maximum load capacity of the delivery vehicle, and cannot be served by multiple delivery vehicles or the same vehicle multiple times.
8. Delivery vehicles can only serve one delivery route.

3.2. Objective function. The optimization objective of the vehicle routing problem with time windows is to select a reasonable driving route while satisfying capacity and time window constraints, in order to minimize the total cost of the entire delivery process [19]. Based on this, the objective function *Cost* designed in this article consists of three parts, namely transportation cost D, capacity penalty cost W, and time penalty cost P. The specific definition is shown in formula (3.1).

$$Min\ Cost = D + W + P \tag{3.1}$$

The calculations for each section are shown below.

1) Transportation cost D**.** The transportation distance affects the transportation cost, which in turn affects the total cost of logistics distribution. The transportation cost in this article is represented by the transportation distance, which is the total distance traveled by all participating delivery vehicles. The calculation of D is shown in formulas 3.2 and 3.3.

$$D = \sum_{k=1}^{K} \sum_{i=0}^{N} \sum_{j=0}^{N} d_{ij} * x_{ij}^{k} \tag{3.2}$$

$$x_{ij}^{k} = \begin{cases} 1 & \text{if the delivery vehicle } k \text{ goes from customer point } i \text{ to customer point } j \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Among them, K is the number of vehicles participating in the delivery task; N is the total number of customer points; d_{ij} is the Euclidean distance between customer point i and customer point j; x_{ij}^{k} is a decision variable that indicates whether delivery vehicle k is from customer point i to customer point j; If so, x_{ij}^{k} is 1, otherwise it is 0.

2) Capacity penalty cost W**.** Delivery vehicles depart from the distribution center and deliver goods to customer points in sequence. After completing the delivery task, they return to the distribution center to form a distribution route. If the total cargo capacity of the delivery vehicle on the delivery route exceeds the maximum load capacity of the vehicle, which violates the capacity constraint, there will be a capacity penalty cost. The calculation of W is shown in formulas 4 to 6.

$$C_k = \sum_{i=1}^{l} q_i \tag{3.4}$$

$$w_k = \alpha * max(C_k - Q, 0) \tag{3.5}$$

$$W = \sum_{k=1}^{K} w_k \tag{3.6}$$

Among them, q_i represents the demand for goods at customer point i; l is the length of the delivery route, which refers to the number of customer points that the vehicle passes through; Q is the maximum load capacity of the vehicle; α is the penalty coefficient for violating capacity constraints; k is the number of vehicles participating in the delivery task; C_k is the sum of the demand for customer points that vehicle k passes through; w_K is the penalty cost for vehicle k violating capacity constraints; W is the penalty cost for the capacity of all delivery vehicles.

3) Time penalty cost P**.** This article combines the actual situation of logistics distribution and studies the vehicle routing problem with soft time windows. Delivery vehicles are required to deliver goods to customer points within the specified time frame, and there will be no time penalty costs incurred. If the logistics delivery vehicle fails to deliver the goods to the customer's location within the specified time window, there will be a certain time penalty cost. Based on the established problem model, while considering the complexity and operational efficiency of the algorithm, this paper only penalizes late arriving vehicles and calculates the time penalty cost. The calculation of P is shown in formulas (3.7) to (3.9).

$$P_i^k\left(S_i\right) = \beta * \max(S_i - L_i, 0) \tag{3.7}$$

$$P_0^k\left(B_0\right) = \beta * \max(B_0 - L_0, 0) \tag{3.8}$$

$$P = \sum_{k=1}^{K}\sum_{i=1}^{N} P_i^k\left(S_i\right) + \sum_{k=1}^{K} P_0^k\left(B_0\right) \tag{3.9}$$

Among them, β Is the penalty coefficient for violating the time window constraint; S_i is the actual start time of service for customer point i; L_i is the right time window for customer point i, which is the latest service start time for customer point i; $P_i^k\left(S_i\right)$ is the penalty cost for delivery vehicle k violating the time window constraint at customer point i; $P_0^k\left(B_0\right)$ is the penalty cost for violating the end time of the distribution center time window after vehicle k returns to the distribution center; B_0 is the time when the vehicle returns to the distribution center; L_0 is the end time of the distribution center's time window, which is the latest time the vehicle returns to the distribution center; P is the penalty cost for all delivery vehicles violating time window constraints at all customer points and returning to the distribution center.

The calculation of S_i and B_0 is shown in formulas (3.10) to (3.12).

$$S_i = Max(R_i, E_i) \tag{3.10}$$

$$R_i = S_{i-1} + T_{i-1} + d_{(i-1,i)} \tag{3.11}$$

$$B_0 = S_{end} + T_{end} + d_{(end,0)} \tag{3.12}$$

Among them, R_i is the time when the vehicle arrives at customer point i; E_i is the left time window of customer point i, which is the earliest service start time of customer point i; S_{i-1} is the actual service start time of the previous customer at customer point i; T_{i-1} is the service time required by the previous customer at customer point i; $d_{(i-1,i)}$ is the distance between customer i and the previous customer; S_{end} is the start time of service for the last customer on the delivery route; T_{end} is the service time required for the last customer; $d_{(end,0)}$ is the distance between the last customer and the distribution center.

4. Design of Improved Genetic Algorithm for VRPTW.

4.1. Chromosome coding.

4.1.1. Encoding method. Due to the fact that the quality of encoding directly affects the efficiency and results of algorithm operations, the primary issue in implementing genetic algorithms is to choose the appropriate chromosome encoding method. This article focuses on the characteristics of vehicle routing problems with time windows and adopts natural number encoding. The specific encoding method is as follows.

Number N delivery customer points sequentially, using $1, 2, , N$. M is the predetermined maximum number of vehicles used. K is the actual number of vehicles participating in delivery. A chromosome encoded using natural numbers is shown in Fig. 4.1. R_i is the i-th customer point. Chromosome length is $N + M$.

4.1.2. Conversion between chromosome and delivery route. Find the position of the vehicle number in the chromosome, i.e. Chrome> N, and extract the route before the vehicle number, which is the corresponding customer point route that the vehicle passes through. The driving routes of all delivery vehicles constitute a delivery plan.

For example, assuming that the number of customer points N is 7, the maximum number of vehicles used M is 10, and the actual number of vehicles participating in delivery K is 3, a chromosome is shown in Fig. 4.2.

The delivery plan is as follows:

Delivery route for the first vehicle: Route $\{1\} = \{3, 5, 6\}$;

Delivery route for the second vehicle: Route $\{2\} = \{1, 2\}$;

Delivery route for the third vehicle: Route $\{3\} = \{4, 7\}$.

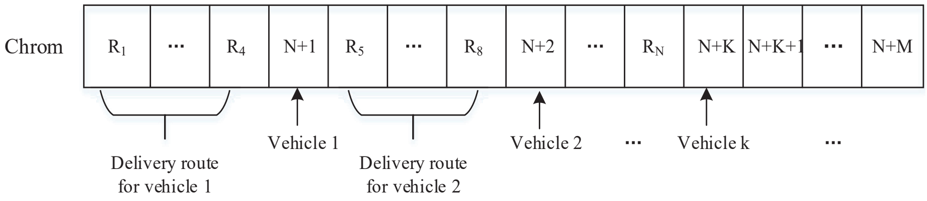


Fig. 4.1: Chromosome coding diagram

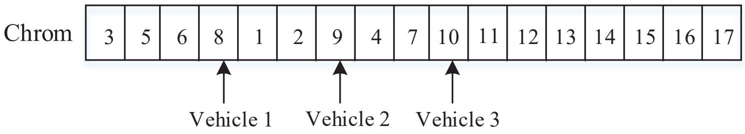


Fig. 4.2: A hypothetical chromosome diagram

4.2. Initialize population. The initial population has a direct impact on the solution of vehicle routing problems with time windows. If the initial population is randomly generated, the generated chromosomes may not be excellent enough, which will have a certain impact on the optimization speed and quality of the optimal solution of the algorithm, leading to the inability of the algorithm to seek the global optimal solution [20]. Therefore, this article designs a heuristic initialization algorithm to generate a better initial population.

The specific steps for initializing the algorithm are as follows.

Step 1. Establish a sequence of traversing customer points The method is as follows:

First, randomly select a customer point i from N customer points;

Then, follow the following sequence seq to traverse each customer point;

If $i = 1$, $seq = [1 : n]$

If $i = n$, $seq = [n, 1 : i - 1]$;

Otherwise, $seq = [i : n, 1 : i - 1]$.

Step 2. Start traversing to obtain the vehicle delivery route

The method is as follows:

First, create a cell array called $Route = cell(k, 1)$ to store the vehicle delivery route. The initial value of k is 1, and $Route\{k\}$ stores the customer point route passed by the kth vehicle.

Then, according to the seq sequence, add the customer points to $Route\{k\}$ in sequence.

When adding a customer point, it is necessary to determine whether the total demand for goods at the customer point Rout $\{k\}$ after addition exceeds the maximum load capacity of the vehicle. If not, add it; otherwise, it cannot be added and an additional vehicle needs to be added to store the customer point.

Finally, arrange the customer points in $Route\{k\}$ in ascending order of the left time window values.

Step 3. Generate initial population

Convert the delivery routes obtained in step 2 into chromosome and generate the initial population.

4.3. Determine fitness function. In genetic algorithms, the fitness function value is used to evaluate the quality of chromosomes and is also the basis for individual evolution. If the fitness value of a chromosome is higher, it indicates a higher degree of excellence of the chromosome [6]. There is a high probability that the individual will be replicated to the next generation. The fitness function value of chromosomes is generally related to the objective function value. In the model constructed in this article, the objective function is to minimize the total logistics cost. Therefore, the fitness function value adopts the reciprocal of the objective function value. The smaller the objective function value, the larger the fitness function value. The definition of fitness function is shown in formula (4.1).

$$Fitness_i = \frac{1}{Cost_i}, \quad i = (1, 2, ..., Z) \tag{4.1}$$

Among them, $Cost_i$ is the objective function value of the i-th chromosome, and the specific calculation is shown in 3.2; $Fitness_i$ is the fitness function value of the i-th chromosome; Z represents the population size.

4.4. Selection operation. The selection operation is based on the fitness value of chromosomes to determine whether they can enter the next generation. The higher the fitness value, the higher the probability of entering the next generation population. On the contrary, the smaller the fitness value, the less likely it is to enter the next generation. This article selects some individuals from the population according to the set generation gap. The selection method adopts random traversal sampling, which generates multiple equally spaced marker pointer positions at once to select the corresponding individuals.

The specific steps for selecting operation are as follows.

Step 1. Calculate the spacing P between pointers, as shown in formulas (4.2) and (4.3);

$$P = \frac{\sum_{i=1}^{Z} Fitness_i}{N} \tag{4.2}$$

$$N = Gap * Z \tag{4.3}$$

Among them, Z is the population size, *Gap* is the generation gap, and N is the number of individuals to be selected.

Step 2. Randomly generate the starting point pointer position, denoted as *start*, which is a random number between 0 and P;

Step 3. Calculate the positions of each *pointer*, denoted as pointers, as shown in formula (4.4);

$$pointers = start + i * P, \quad (i = 0, 1, \cdots, N-1) \tag{4.4}$$

Step 4. Select N individuals based on the positions of each pointer.

4.5. Cross operation. Cross operation refers to the operation of exchanging one or more bits between two parent individuals to generate a new individual. Cross operation is an important operation in searching the solution space, which can search for the optimal solution within the maximum range. It not only affects the computational efficiency of genetic algorithms, but also affects the computational results of genetic algorithms. This article performs crossover operations on parent individuals based on a certain probability of crossover.

The specific steps for cross operation are as follows.

Step 1. Generate a random number *rand* between 0 and 1. If $rand <= P_c$, P_c is the crossover probability, then go to Step 2 and perform a crossover operation on two adjacent individuals of the parent;

Step 2. Randomly generate two intersection points, denoted as r_1 and r_2, respectively. Swap the two sets of gene sequences between r_1 and r_2, placing them at the forefront of the corresponding chromosome;

Step 3. Eliminate duplicate gene loci to obtain the final offspring chromosome.

For example, A and B are a pair of chromosomes in the parent generation, A' and B' are the offspring chromosomes obtained after crossover operation, as shown in Fig. 4.3.

4.6. Mutation operation. Mutation operator refers to a change in the gene value of one or a few positions in a chromosome, which transforms into other alleles and generates a new individual. The mutation operator can fine tune the new individuals generated by the crossover operator, thereby improving the algorithm's local search ability.

The specific steps for mutation operation are as follows.

Step 1. Generate a random number *rand* between 0 and 1. If $rand <= P_m$, P_m is the mutation probability, then go to Step 2 and perform a mutation operation on the parent chromosome;

Step 2. Randomly generate two gene variants and swap the genes at the two variant positions.

For example, C is the parent chromosome. C' is he offspring chromosome obtained after mutation operation, as shown in Fig. 4.4.

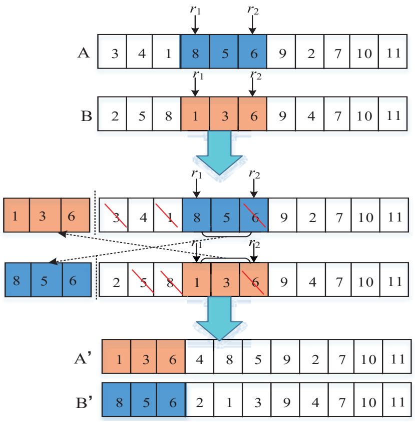

r1
r2
A 3 4 1 8 5 6 9 2 7 10 11
B 2 5 8 1 3 6 9 4 7 10 11
1 3 6
3 4 1 8 5 6 9 2 7 10 11
8 5 6
2 5 8 1 3 6 9 4 7 10 11
A' 1 3 6 4 8 5 9 2 7 10 11
B' 8 5 6 2 1 3 9 4 7 10 11

Fig. 4.3: Example of cross operation

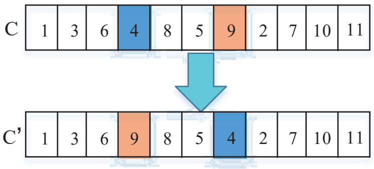


Fig. 4.4: Example of mutation operation

4.7. Local search operation. After selection, crossover, and mutation operations, the diversity of the population is increased, and the algorithm can avoid premature convergence. However, the quality of the optimal solution obtained by the algorithm is not high [11]. In order to improve the optimization and local search capabilities of the algorithm, this paper constructs a local search operation, which mainly includes creating local correlation group and reconstructing vehicle path.

1) Creating local correlation group. Create a local correlation group D based on the correlation between customer points. The algorithm for creating D is described as follows.

Step 1. Randomly select a customer point from the original customer point set and move it to D;

Step 2. Calculate the correlation between the customer point and the other customer points, as shown in formulas (4.5) to (4.7);

$$R_{ij} = \frac{1}{C_{ij} + V_{ij}} \tag{4.5}$$

$$C_{ij} = \frac{d_{ij}}{md_i} \tag{4.6}$$

$$V_{ij} = \begin{cases} 0 & \text{if customer point } i \text{ and customer point } j \text{ are on the same vehicle path} \\ 1 & \text{otherwise} \end{cases} \tag{4.7}$$

Among them, d_{ij} is the distance between customer point i and customer point j; mdi is the maximum distance between customer point i and other customer points.

From the above formulas, it can be seen that the R_{ij} value depends on d_{ij} and V_{ij}. If $V_{ij} = 0$, d_{ij} is smaller and R_{ij} is larger.

Step 3. Select the customer point with the highest relevance and move it to D;

Step 4. If the population size of D is not satisfied, then go to Step 2; otherwise, the algorithm stops.

2) Reconstructing vehicle path. The customer points in the local correlation group D need to be reinserted back into the vehicle path to obtain a new vehicle delivery plan. In order to improve the optimization ability and solution quality, this article ensures that the inserted vehicle path satisfies capacity constraints and time window constraints when inserting customer points back into the vehicle path. Based on this, we construct a reinsertion heuristic algorithm to find the optimal insertion vehicle and location, and reconstruct the vehicle path. The description of the reinsertion heuristic algorithm is as follows.

Step 1. For a certain customer point i in D, first determine whether it can be added to the path $r(r_1, r_2, ..., r_l)$ passed by vehicle k, that is, first determine whether it meets the capacity constraint, and the judgment method is as shown in formula (4.8). If the capacity constraint is not met, customer point i cannot be inserted into the vehicle path $r(r_1, r_2, ..., r_l)$. If the capacity constraint is met, go to Step 2;

$$\sum_{a=1}^{a=l} q\left(r_a\right) + q(i) \leq Q \tag{4.8}$$

Among them, $\sum_{a=1}^{a=l} q\left(r_a\right)$ is the sum of the demand for goods from all customer points in path $r(r_1, r_2, ..., r_l)$; $q(i)$ is the demand for goods at customer point i; Q is the maximum load capacity of the vehicle.

Step 2. Insert customer point i into path $r(r_1, r_2, ..., r_l)$. There will be $l + 1$ insertion point positions and generate $l + 1$ new paths. Determine whether all customer points on the new path and vehicle returning to the distribution center meet the time window constraint, as shown in formula (4.9).

$$S_j \leq L_j,\ B_0 \leq L_0 \tag{4.9}$$

Among them, j is the customer point on the new path; S_j and B_0 are the actual start time of service at customer point j and the time when the vehicle returns to the distribution center, and the specific calculations are shown in formulas (3.10) to (3.12); L_j and L_0 are the end times of the time windows for customer point j and distribution center.

Step 3. If the time window constraint is met, record the insertion point position, vehicle number, and distance increment.

Step 4. Find the insertion point position and vehicle number with the smallest distance increment, which is the optimal insertion vehicle and position. If it exists, insert customer point i and reconstruct the vehicle path; otherwise, add a new vehicle and deliver to customer point i.

4.8. Reconstruct a new population and save the current optimal solution. The new population consists of elite individuals from the parent generation and offspring. Descendants are obtained through selection, crossover, mutation, and local search operations on their parents. Parent elite individuals are obtained by performing elite preservation operations on the parent population, that is, individuals with smaller objective function values selected in proportion to (1-*Gap*).

Calculate the objective function value for the new population, select the chromosome with the smallest objective function value, and save it as the current optimal solution.

4.9. Calculate the total distance and record the number of paths that violate constraints. Based on the optimal solution, obtain the delivery plan and calculate the total distance traveled by all delivery vehicles. Determine whether each vehicle path violates the load capacity constraint or time window constraint. If it does, record the number of vehicle paths that violate constraints.

4.10. Algorithm framework. The algorithm framework proposed in this article is as follows.

Step 1: Initialize parameters;
Step 2: Initialize the population;
Step 3: Calculate the fitness value of each chromosome in the population;
Step 4: Perform selection operation;
Step 5: Perform cross operation;
Step 6: Perform mutation operation;
Step 7: Perform local search operation;
Step 8: Reconstruct a new population and save the optimal solution;
Step 9: Calculate the total distance and record the number of paths that violate constraints;
Step 10: If the algorithm does not meet the termination condition, go to Step 3.

5. Simulation Experiments. In order to effectively verify the feasibility and effectiveness of the proposed improved genetic algorithm (IGA) in solving VRPTW problems, this paper conducts comparative experiments with the genetic algorithm without local search operation(GA) and the hybrid genetic algorithm (HGA) improved by others [3].

Table 5.1: Parameter Settings

Parameter	*Value*
Population size (Z)	100
maximum iterations	100
Generation gap (Gap)	0.9
crossover probability (P_c)	0.9
mutation probability (P_m)	0.05
The size of local correlation group D	15
Penalty coefficient for violating capacity constraint (α)	10
Penalty coefficient for violating time window constraints (β)	100

Table 5.2: Statistical Results of C101

Algorithm	*GA*	*HGA*	*IGA*
Number of vehicles used	22	11	10
The total distance traveled by vehicles	3557.7041	904.2931	828.9369
Number of paths that violate constraints	12	0	0
The number of times the optimal solution found in 10 runs	3	7	10
The iteration number of the earliest discovered optimal solution in 10 runs	72	68	41

Table 5.3: Statistical Results of C201

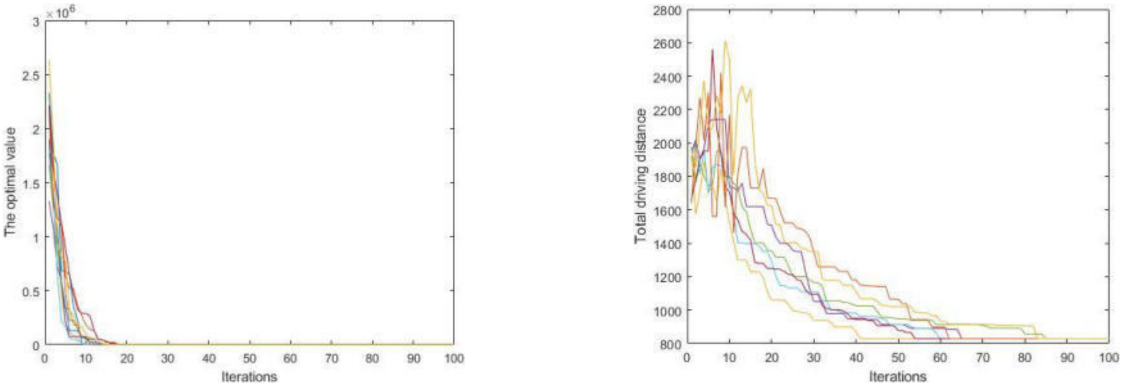
Algorithm	*GA*	*HGA*	*IGA*
Number of vehicles used	21	4	3
The total distance traveled by vehicles	3395.7609	621.5892	591.5566
Number of paths that violate constraints	15	0	0
The number of times the optimal solution found in 10 runs	4	8	10
The iteration number of the earliest discovered optimal solution in 10 runs	8	39	25

5.1. Test examples. The C-type dataset in the Solomon dataset [15] represents that the location of customer nodes is generated based on a structural distribution, divided into two different series of data, namely C1 and C2. We selected two instances, C101 and C201, as representatives of the C-type dataset to analyze the efficiency of our algorithm in solving this type of problem. The calculation example contains some known information, such as the maximum number of vehicles used, the maximum load capacity of vehicles, the location coordinates and time windows of 100 customer points, and the demand for goods.

5.2. Experimental environment and parameter settings. The computer configuration used in the simulation experiment is: Core dual core CPU, 2.50GHz, 32GB memory, and Windows 10 system. The simulation software used is MATLAB R2021b. The parameter settings in the algorithm are shown in Table 5.1.

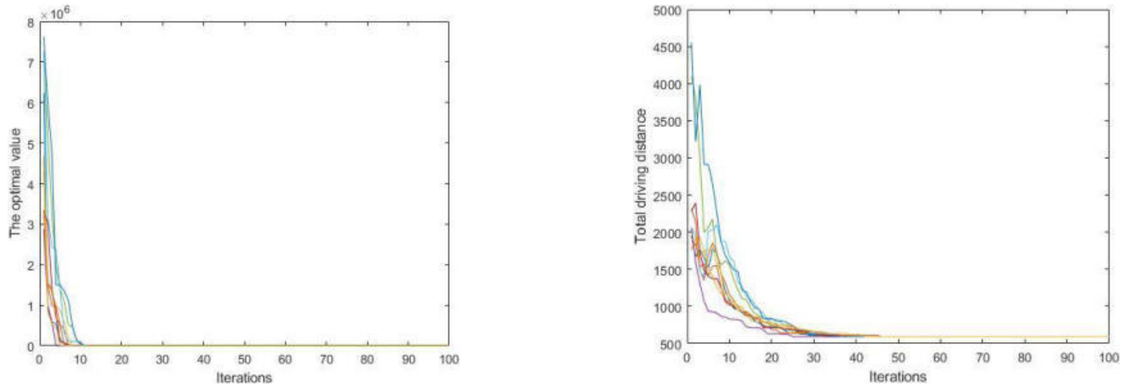
5.3. Experimental results and comparative analysis. To avoid the influence of randomness on the algorithm, we run GA, HGA, and IGA algorithms 10 times on each of the two examples, and recorded the relevant experimental data of the obtained optimal solution. The statistical results are shown in Table 5.2 and Table 5.3.

In order to better observe the evolution process of the proposed algorithm (IGA) and more intuitively reflect its performance, we draw an evolution graph of the optimal value and total distance obtained in each



(a) Evolution diagram of the optimal value (b) Evolution diagram of total driving distance

Fig. 5.1: Evolution diagram of optimal solution related data for C101



(a) Evolution diagram of the optimal value (b) Evolution diagram of total driving distance

Fig. 5.2: Evolution diagram of optimal solution related data for C201

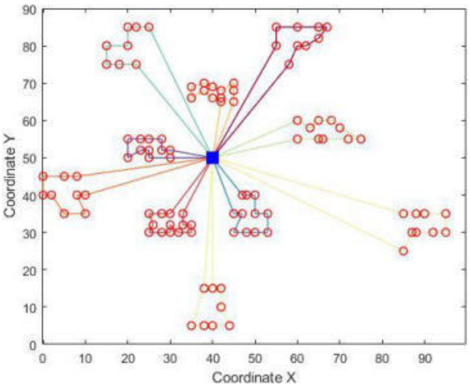
iteration. Taking into account the impact of randomness on algorithm performance, we superimpose the results of 10 runs to evaluate algorithm performance more realistically and objectively. The experimental results are shown in Fig. 5.1 and Fig. 5.2.

From Table 5.2 and Fig. 5.1, it can be seen that the IGA algorithm can find the currently known optimal total distance of C101, which is 828.9369, in each of the 10 runs. At the earliest, it can be found in 41 iterations. From Table 5.3 and Fig. 5.2, it can be seen that the IGA algorithm can find the currently known optimal total distance of C201, which is 591.5566, in each of the 10 runs. At the earliest, it can be found in 25 iterations. At the same time, it can be clearly seen from Fig. 5.1 and Fig. 5.2 that in the early stages of evolution, the total delivery cost and total driving distance of C101 and C201 both show a significant decrease, which also reflects the good optimization ability of the proposed algorithm in solving vehicle routing problems with time windows.

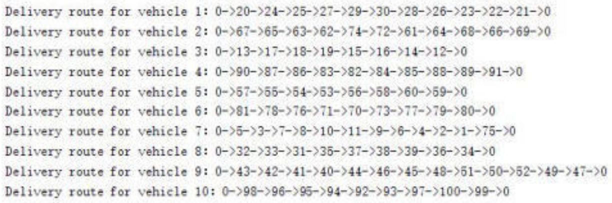
The optimal delivery plan for C101 is shown in Fig. 5.3, requiring 10 delivery vehicles.

The optimal delivery plan for C201 is shown in Fig. 5.4, requiring 3 delivery vehicles.

In order to provide a more intuitive comparison of algorithm performance, we draw an evolutionary comparison chart of the total driving distance obtained by the three algorithms in a random experiment, as shown in Fig. 5.5 for C101 and Fig. 5.6 for C201.



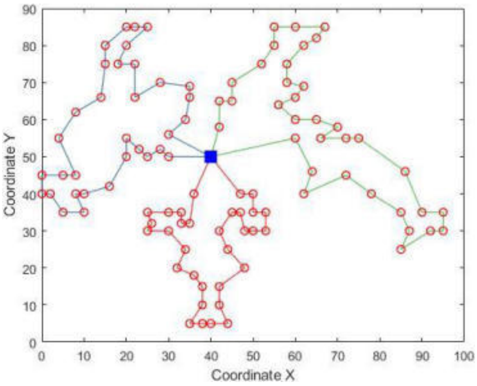
(a) Delivery route map



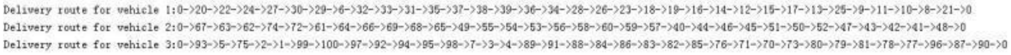
```
Delivery route for vehicle 1: 0->20->24->25->27->29->30->28->26->23->22->21->0
Delivery route for vehicle 2: 0->67->65->63->62->74->72->61->64->68->66->69->0
Delivery route for vehicle 3: 0->13->17->18->19->15->16->14->12->0
Delivery route for vehicle 4: 0->90->87->86->83->82->84->85->88->89->91->0
Delivery route for vehicle 5: 0->57->55->54->53->56->58->60->59->0
Delivery route for vehicle 6: 0->81->78->76->71->70->73->77->79->80->0
Delivery route for vehicle 7: 0->5->3->7->8->10->11->9->6->4->2->1->75->0
Delivery route for vehicle 8: 0->32->33->31->35->37->38->39->36->34->0
Delivery route for vehicle 9: 0->43->42->41->40->44->46->45->48->51->50->52->49->47->0
Delivery route for vehicle 10: 0->98->96->95->94->92->93->97->100->99->0
```

(b) Specific delivery route map

Fig. 5.3: The optimal distribution plan for C101



(a) Delivery route map



```
Delivery route for vehicle 1:0->20->22->24->27->30->29->6->32->33->31->35->37->38->39->36->34->28->26->23->18->19->16->14->12->15->17->13->25->9->11->10->8->21->0
Delivery route for vehicle 2:0->67->63->62->74->72->61->64->66->69->68->65->49->55->54->53->56->58->60->59->57->40->44->46->45->51->50->52->47->43->42->41->48->0
Delivery route for vehicle 3:0->93->5->75->2->1->99->100->97->92->94->95->98->7->3->4->89->91->88->84->86->83->82->85->76->71->70->73->80->79->81->78->77->96->87->90->0
```

(b) Specific delivery route map

Fig. 5.4: The optimal distribution plan for C201

It can be clearly seen from Fig. 5.5 and Fig. 5.6 that the genetic algorithm without local search operation (GA) cannot converge to the global optimal solution and has poor performance, while the improved genetic algorithm with local search operation (IGA) shows the best performance. The performance of HGA is also inferior to that of IGA. This further confirms the effectiveness of the local search operation in the algorithm proposed in this paper. It not only avoids the defect of genetic algorithms easily falling into local optima, but also enhances algorithm diversity and obtains better optimal solutions.

The above experiment is a specific analysis of the C101 and C201 examples in the Solomon dataset, and has achieved good results. These results indicate that the improved genetic algorithm in this paper is significantly better than GA and HGA in solving quality and stability issues. It is a feasible and effective method for solving vehicle routing problems with time windows.

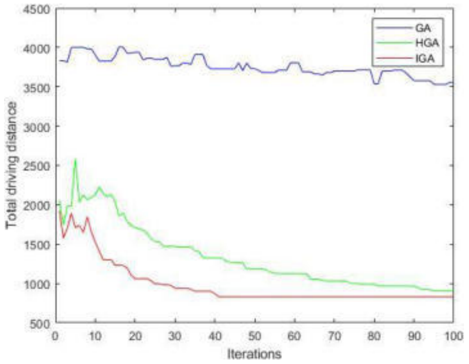


Fig. 5.5: Comparison chart of C101 total driving distance

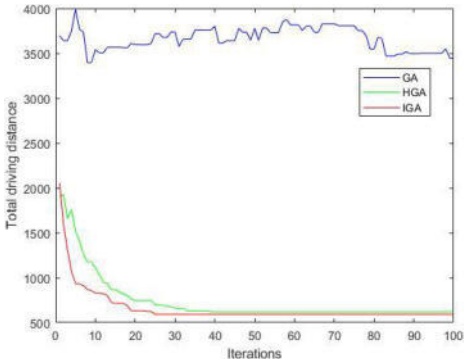


Fig. 5.6: Comparison chart of C201 total driving distance

6. Conclusions and Further Work. This article first analyzes the vehicle routing problem with time windows, constructs the objective function of the problem, and establishes a mathematical model for the problem. In order to improve the optimal solution quality of genetic algorithm in solving vehicle routing problems with time windows, this paper proposes an improved genetic algorithm. The main improvements are as follows: (1) A heuristic initialization algorithm is designed to generate a high-quality initial population; (2) In order to enhance the diversity of the population and avoid premature convergence of the algorithm, selection, crossover, and mutation operations are designed; (3) A local search operation is designed to improve the optimization and local search capabilities of genetic algorithms. Finally, simulation experiments have verified that the improved genetic algorithm is a feasible and effective method for solving vehicle routing problems with time windows. Our future work will be to combine other intelligent optimization algorithms for more in-depth theoretical analysis and algorithm research. At the same time, we will further expand our application scope and conduct research on other types of vehicle routing problems.

Acknowledgments. This work was supported by Anhui University Natural Science Research Key Project under Grant (No.KJ2020A0539, No.2022AH051325, No.2024AH051466).

REFERENCES

[1] H. T. T. Ai, N. T. Thi, and N. V. Can, *A multiple objective model for vehicle routing problem with time windows: a case study*, Applied Mechanics and Materials, 889 (2019), pp. 588–596.

[2] C. Altman, G. Desaulniers, and F. Errico, *The fragility-constrained vehicle routing problem with time windows*, Transportation Science, 57 (2023), pp. 552–572.

[3] M. A. Ara, M. T. Ahmed, and N. Yeasmin, *Optimisation model for simultaneous delivery and pickup vehicle routing problem with time windows*, International Journal of Services and Operations Management, 43 (2022), pp. 145–168.

[4] Y. Chen, P. Huang, J. Chen, Z. Wang, Y. Shen, and X. Fan, *Hybrid particle swarm optimization algorithm based on hierarchical learning and different evolution for solving capacitated vehicle routing problem*, Computer Science, 49 (2022), p. 7.

[5] L. M. Dalbah, M. A. Al-Betar, M. A. Awadallah, and R. A. Zitar, *A modified coronavirus herd immunity optimizer for capacitated vehicle routing problem*, Journal of King Saud University-Computer and Information Sciences, 34 (2022), pp. 4782–4795.

[6] C. S. Ganesh, R. Sivakumar, and N. Rajkumar, *Retraction note: Soft computing-based fuzzy time series model for dynamic vehicle routing problem*, 2023.

[7] M. He, Z. Wei, X. Wu, and Y. Peng, *An adaptive variable neighborhood search ant colony algorithm for vehicle routing problem with soft time windows*, IEEE Access, 9 (2021), pp. 21258–21266.

[8] H.-w. JIANG, T. GUO, and Z. YANG, *Research progress of vehicle routing problem*, ACTA ELECTONICA SINICA, 50 (2022), p. 480.

[9] T. S. Khoo and B. B. Mohammad, *The parallelization of a two-phase distributed hybrid ruin-and-recreate genetic algorithm for solving multi-objective vehicle routing problem with time windows*, Expert Systems with Applications, 168 (2021), p. 114408.

[10] Y. Liu, Y. Yu, Y. Zhang, R. Baldacci, J. Tang, X. Luo, and W. Sun, *Branch-cut-and-price for the time-dependent green vehicle routing problem with time windows*, INFORMS Journal on Computing, 35 (2023), pp. 14–30.

[11] S. Muñoz-Herrera and K. Suchan, *Local optima network analysis of multi-attribute vehicle routing problems*, Mathematics, 10 (2022), p. 4644.

[12] D. A. Neira, M. M. Aguayo, R. De la Fuente, and M. A. Klapp, *New compact integer programming formulations for the multi-trip vehicle routing problem with time windows*, Computers & Industrial Engineering, 144 (2020), p. 106399.

[13] B. Pan, Z. Zhang, and A. Lim, *A hybrid algorithm for time-dependent vehicle routing problem with time windows*, Computers & Operations Research, 128 (2021), p. 105193.

[14] M. Roboredo, R. Sadykov, and E. Uchoa, *Solving vehicle routing problems with intermediate stops using vrpsolver models*, Networks, 81 (2023), pp. 399–416.

[15] M. M. Solomon, *Algorithms for the vehicle routing and scheduling problems with time window constraints*, Operations research, 35 (1987), pp. 254–265.

[16] Y. Wang, S. Luo, J. Fan, M. Xu, and H. Wang, *Compensation and profit allocation for collaborative multicenter vehicle routing problems with time windows*, Expert Systems with Applications, 233 (2023), p. 120988.

[17] Y. Wang, Y. Wei, X. Wang, Z. Wang, and H. Wang, *A clustering-based extended genetic algorithm for the multidepot vehicle routing problem with time windows and three-dimensional loading constraints*, Applied Soft Computing, 133 (2023), p. 109922.

[18] H. Wu, Y. Gao, W. Wang, and Z. Zhang, *A hybrid ant colony algorithm based on multiple strategies for the vehicle routing problem with time windows*, Complex & intelligent systems, (2021), pp. 1–18.

[19] Y. Xiang, Y. Zhou, H. Huang, and Q. Luo, *An improved chimp-inspired optimization algorithm for large-scale spherical vehicle routing problem with time windows*, Biomimetics, 7 (2022), p. 241.

[20] N. Yu, B. Qian, R. Hu, Y. Chen, and L. Wang, *Solving open vehicle problem with time window by hybrid column generation algorithm*, Journal of Systems Engineering and Electronics, 33 (2022), pp. 997–1009.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Jan 3, 2024
Accepted: Sep 3, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 136–150, DOI 10.12694/scpe.v26i1.3537

RECOVERY MODELING AND ROBUSTNESS STUDY AFTER CASCADING FAILURES IN LOGISTICS-BASED NETWORKS

XIAODONG QIAN* AND SICHEN WANG†

Abstract. In order to ensure the normal operation of the logistics network and improve the robustness of the network in case of cascading failure faults, this study introduces the concept of node recovery threshold based on the existing failure model to optimize the recovery time lag. Further, a criticality-first recovery model is proposed, which defines the capacity of a recovery node as a function related to its original capacity and opens the recovery node selectively to critical neighboring nodes to reduce the risk of secondary failure. Finally, a postal logistics network in Northwest China is used as a case study to investigate the recovery robustness of this network when it encounters cascading failures. The effects of various parameter variations on the network robustness are examined through experimental simulations. The experimental results show that timely recovery measures can significantly reduce the number of failed nodes when cascade failure occurs in the logistics network; setting a higher recovery threshold can reduce the impact of cascade failure on the network, effectively reduce the scale of network failure, and thus significantly improve the robustness of the logistics network; at the same time, increasing the capacity parameter can effectively delay the time of cascade failure in the network, and can slightly improve the robustness of the network.

Key words: complex networks, logistics networks, robustness, cascading failures, recovery modeling

1. Introductory.

With the tremendous growth of science and technology, and economic development, the logistics industry in its context also began to flourish, and logistics facilities continued to improve, the growing development of transportation, greatly improved the operating conditions of logistics, making the traditional logistics industry the transformation of the network, and a highly efficient and reliable logistics network is a prerequisite for the healthy development of the logistics industry and the foundation. However, in recent years, all kinds of emergencies have occurred frequently, such as the general strike of German freight train drivers in 2015, the outbreak of the new crown epidemic in 2020, the congestion of the Suez Canal caused by the stranding of the cargo ship ”Chang Ci” in 2021, and the extremely heavy rain in Henan in 2022, which had a serious impact on the normal operation of the logistics network and disrupted the normal production of social life. These events have seriously affected The logistics network operates normally, disturbed the normal order of production and life of the society, and threatened the peace and stability of the countries in the world. In real life, The efficient and dependable movement of goods is dependent on the proper operation of the logistics network, and all types of emergencies are unavoidable; thus, effective and reliable preventive measures are critical to the overall safety of the logistics network.

Numerous scholars have found that there is no one hundred percent reliable real network, and the failure of the network is unavoidable, and the same is true for the logistics network, through the study of the structural characteristics, dynamic characteristics reliability, and other characteristics of many real networks. At present, some scholars optimize the network by changing the topology of the network to improve the robustness of the network, and this scheme can effectively improve the robustness of the network, but in practice, it may not be easy to change the original topology of the network due to the limitations of various aspects, and we can’t improve the robustness of the network promptly by adjusting and improving the topology of the network. Therefore, the recovery characteristics of the network have aroused the research interest of the majority of scholars. In real logistics networks, the network generally has a certain ability to resist risk and recovery, and when the logistics network fails, the failed nodes will be restored to normal nodes through some recovery strategies. It can assist us in increasing the robustness of the logistics network through recovery mechanisms

[*] School of Economics and Management, Lanzhou Jiaotong University, Lanzhou 730070, China; School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (qianxd@mail.lzjtu.cn).

[†] School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China.

without affecting the network topology. avoiding policy, technical and economic constraints, and reducing costs. Therefore, it has become an important research topic to formulate recovery strategies to improve the robustness of logistics networks when unexpected events occur.

2. Research Basis. In recent years, research based on complex network theory [1, 2, 3, 4, 5, 6] has attracted widespread attention in many fields, and currently, research on complex networks focuses on complex network modeling, network robustness, network recovery, and network optimization.

In terms of constructing a logistics network, Qin et al [7] built a network using epidemiological modeling and network coupling, and argued that the peak infection rate is an indicator of the infectious disease in a certain location, regardless of network topology. Yang et al [8] constructed an evolutionary model of supply chain network enterprise cooperation based on a bipartite power law distribution in a complex network environment and proved that the degree of adaptation is the dominant factor in the generation of bipartite power law distribution. Chen et al [9] investigated the resilience of the logistics network to node failure under the background of the logistics industry disruption, simulated and proposed algorithms for two cascading failure scenarios, proving the improvement of resilience by different adjustment strategies. Wang et al [10] constructed a multi-stage three-level logistics network model for the dynamic logistics network optimization problem and proposed a dynamic adaptive multi-objective differential evolutionary algorithm to solve the model, proving that the algorithm can compute the optimal feasible supply solutions for each stage of the logistics network. Zhang et al [11] proposed a multi-modal express logistics network optimization decision-making model from the perspective of low carbon economy and proposed a corresponding optimization solution algorithm by analyzing the topology of the logistics network and comparing and analyzing the advantages and disadvantages of different transportation modes.

In terms of network robustness, Qian et al [12] proposed a diffusion model for risk factors in logistics networks based on the contagion model in complex networks, using node topological weights and supply and demand fluctuations to improve the mechanism of diffusion. Wang et al [13] introduced dynamic factors on the basis of the original initial residual capacity load redistribution strategy, and by adjusting the network cost e and capacity parameter gamma, the logistics cost is effectively reduced and the controllable robustness against cascading failures is enhanced. Zhen et al [14] built a cascade failure model based on Dynamic Control of Node Redundancy Capacity (DRC) by introducing a network phase change critical factor μ based on the literature [13].The probability of cascade failure triggered by a node failure is measured by defining a phase change critical factor μ in the network,and the correlation μ between the network robustness and μ is analyzed. German R M et al [15] proposed a new method to measure the robustness of inverse logistics networks, using an integer mixed linear programming model to design the network and analyze its robustness, and proved the effectiveness of the method. Yang et al [16] combine the bus network and subway network as a multi-subnet composite complex network of urban public transportation to establish the passenger flow transfer rules under node and edge failure. Yu et al [17] proposed a cascading failure model of dependent networks considering dependent edge loads based on literature [16], and studied the robustness of dependent networks by analyzing different load redistribution strategies, different network coupling methods and network attacks.

In terms of network recovery, Zhang & Du [18] developed a geographical The data network model that combines complex networks and hypergraphs. When the network is subjected to an unavoidable attack, a central node recovery technique based on global and community data is presented to restore network performance with the fewest amount of components.Yang et al [19] proposed two supply chain recovery strategies, the suppliers' pre-set emergency inventory strategy and the manufacturer's product change strategy, and demonstrated that the proposed disruption recovery strategy can not only effectively assist suppliers in stopping losses, but also meet market demand. Wu et al [20] proposed a new sequential recovery model, introduced a sequential recovery graph to determine the key nodes and their recovery order, and verified that the network can obtain better performance during the recovery process. Ju et al [21] made improvements on the basis of literature [20] and proposed a system optimal recovery strategy based on the network elastic structure evaluation method. Tang et al [22] proposed a probabilistic recovery model and a stage recovery model and investigated the recovery robustness for four types of networks, ER, WS, NC and BA, and verified that both probabilistic recovery strategy and stage recovery strategy can have an impact on the robustness of the four types of networks. Huang et al [23] constructed an urban subway network recovery model using the average network efficiency as the toughness index and the maximum network resilience as the objective function and proved that the optimal recovery strategy of the urban subway network solved by the genetic algorithm has a higher solving efficiency under the failures of different sizes of stations. Fan et al [24] aimed to develop a realistic cascading failure model for an automotive manufacturing supply chain network to explore its cascading failure characteristics. Based on this model, a two-stage recovery strategy was proposed to enhance the network's cascading failure recovery capability. Yoshihisa et al [25] designed a recovered, efficient reverse logistics network using a cost minimization model with recovery constraints using a Japanese reverse logistics network as an example.

Existing studies have investigated the recovery behavior after the end of network cascade failure and proposed many classical models such as random recovery model, probabilistic recovery model, and local recovery model. The recovery order of the random recovery model is completely random and does not consider the strategic position of nodes or the structural characteristics of the network, which may lead to untimely recovery of key nodes and prolong the overall recovery time of the network, and is suitable for networks with high node homogeneity. The probabilistic recovery model recovers nodes based on preset probability distributions, which can be set based on the importance of nodes, failure probabilities, or other criteria, and is suitable for network environments that require balancing multiple decision factors. Local recovery models perform systematic recovery only in local areas of the network, which may be partitioned based on geographic location or network partitioning, and help to quickly resolve local failures, but may neglect network-wide coherence and efficiency, and are suitable for distributed networks or large geographically dispersed networks, which can quickly deal with local problems without affecting the overall network.

In summary, the existing models do not satisfy the characteristics of logistics network dynamics and complexity. Therefore, in this paper, we construct a node-importance-priority recovery model in which the failure process and the recovery process occur within adjacent time steps, optimize the recovery time lag, redefine the load and capacity of the recovered nodes, and adopt appropriate strategies to reduce the risk of secondary node failures. Compared with the classical model, using the node importance priority recovery model, the recovery order can be arranged according to the importance of the nodes in the network, prioritizing the recovery of those nodes that have the greatest impact on the network function, which can not only recover the core nodes of the network faster, but also significantly improve the overall stability and robustness of the network, which is very in line with the characteristics of the logistics network dynamics and complexity. Finally, an empirical study of the logistics network in Northwest China is carried out in the hope that the study can help repair the damaged network with a more flexible strategy, thus improving the robustness of the logistics network.

3. Research Methodology.

3.1. Analysis of Network Recovery Models under Cascade Failure. In 2002, Motter and Lai et al [26] first introduced the cascading failure mechanism into complex networks and proposed the classical ML capacity-load linear model, which defines the initial load L_i^0 of a node i as a function related to its degree value K_i, namely

$$\begin{cases} L_i^0 = K_i^\theta \\ C_i = (1+\alpha)L_i^0 \end{cases} \tag{3.1}$$

In 2008, literature [27] found that the load and capacity of a real network should have a nonlinear relationship. Therefore, in this paper, we refer to the literature [28],which defines the relationship between the load L_i^0 and capacity C_i of a node as shown in Equation (3.2), and the load is redistributed as shown in Equation (3.3).

$$\begin{cases} L_i^0 = I_i \\ C_i = (1+\beta)\cdot(L_i^0)^\alpha \end{cases} \tag{3.2}$$

where I_i is the importance degree of the node, which is defined as the initial load of the node, the importance degree of the node is obtained by calculating the weight of each index by hierarchical analysis method using the four characteristics of the node, namely, degree value, median, proximity centrality, and eigenvector centrality as the indexes. α and β are the capacity parameters.

$$\begin{cases} \Delta L_{i\to j} = L_i^1 \times \frac{L_j^0}{\Sigma_{m\in\Gamma_i} L_m^0} \\ L_j^1 = \Delta L_{i\to j} + L_j^0 \end{cases} \tag{3.3}$$

Existing studies [29, 30, 31] have shown that the occurrence of cascading failure phenomenon is often accompanied by network recovery, which can improve the robustness of the network without adding redundant nodes and paths to the network. In recent years, scholars have proposed a series of recovery models, which can be applied to different situations and different networks to minimize the loss caused by network failure and improve the robustness of the network as much as possible without changing the network topology.

Currently, the target recovery model based on cascade failure has the following problems: a) The problem of recovery time lag. In reality, complete network failure is only a worst possibility, after the cascade failure phenomenon, the network will keep normal operation of the network through some recovery means, some existing studies start the recovery of nodes only after the complete failure of the network, and do not consider the impact of the introduction of the recovery model on the network's resiliency when cascade failure issues arise. b) Problems of irrational selection of the recovery nodes. Some models recover the network by the failure order or degree value order of the nodes, such models are too one-sided, which may lead to repairing some non-critical nodes and affect the overall recovery of the network, and do not consider the load capacity situation of the recovered nodes as well as the possibility of the recovered nodes having the risk of secondary failures.

3.2. Improvement of Recovery Modeling under Cascade Failure.

3.2.1. Improvement of Node Recovery Time Lag. The original failure model, shown in (3.4), directly defines a node as a failed node when its load exceeds its capacity, and recovery of the failed node begins only after the entire cascade of failed failures is over. The network has been severely affected at this point.

$$\begin{cases} L_j^1 = \Delta L_{i\to j} + L_j^0 < C_j \\ L_j^1 = \Delta L_{i\to j} + L_j^0 > C_j \end{cases} \tag{3.4}$$

Based on the above analysis, when the network cascade failure fault occurs, the faulty nodes need to be recovered in a shorter time, and the recovery of the network nodes should not be started only after the network cascade failure fault is over so that the robustness of the network can be reasonably enhanced. Therefore, based on the original failure model, the concept of node recovery threshold is introduced, When the load on a network's node reaches its capability. but does not exceed the recovery threshold, the node at this point is defined as overloaded rather than directly determined as a node failure, and the network node starts to recover in the process, in which the time required for the faulty node to start recovering needs to be taken into account, and the parameter T is used to characterize the time step required for the node to begin to recovery by using parameter T to characterize the time step required for the node to start recovery, i.e., it takes T time steps for the node to start recovery after the node fails. When a network node fails to recover in time due to various reasons, the node load exceeds the threshold value, and only then the node is determined to be completely failed and removed from the network. The improved failure model shown in (3.5), $(1+\lambda)C_j$ is the recovery threshold of the node and λ is an adjustable parameter.

$$\begin{cases} \Delta L_{i\to j} + L_j^0 \le C_j & ① \\ C_j < \Delta L_{i\to j} + L_j^0 \le (1+\lambda)C_j & ② \\ (1+\lambda)C_j < \Delta L_{i\to j} + L_j^0 & ③ \end{cases} \tag{3.5}$$

In this case, Equation ① indicates that the load distributed by node i to its neighbor node j does not exceed its maximum capacity, and the network is in a normal working state. Equation ② indicates that the load of the node at this moment exceeds its maximum capacity but does not exceed the recovery threshold, the node at this time is defined as an overloaded node, and the load of the overloaded node j will be assigned to its neighboring nodes according to the rules, and the overloaded node recovers in the process. Equation ③ indicates that when the overloaded node fails to recover in time, the load of the node j eventually exceeds the node recovery threshold, and then the node j fails completely.

3.2.2. Network Recovery Model Optimization. In summary, this paper proposes an importance-priority recovery model suitable for real logistics networks, which is in line with the characteristics of unequal importance of nodes in the logistics network, and the recovery of important nodes can help the logistics network to share the pressure of larger goods and ensure the circulation of goods.

The original recovery model is to recover the network through the node's failure sequence or degree value sequence, there are many problems in the selection of the recovery nodes, the most important nodes should be comprehensively selected for recovery, the model uses the node's importance sequence as the basis for recovery, defines the capacity of the recovered node as a function related to the original capacity, and selectively opens up the recovered node to the important nodes around it to reduce the node's secondary failure risk. In this study, when cascading failure occurs, overloaded nodes in the network that exceed the maximum capacity of a node but do not exceed the recovery threshold are defined as recoverable nodes, and these nodes are repaired by the importance-first recovery model. The recovery process of overloaded nodes is defined as follows.

Step 1: Implement a merit recovery strategy according to the importance of the nodes, repair the overloaded nodes in descending order of importance, and the repaired nodes are restored as normal nodes again.

Step 2: Redefine the load of the recovered node, since the node assigns its load to the neighboring nodes when it is overloaded, the load $L_i(R)$ of the recovered node is defined to be zero, That is.

$$L_i(R) = 0 \tag{3.6}$$

Step 3: Redefine the capacity of the restored node, in reality, the restored system components may be more reliable than before the failure [32, 33]. Therefore, define the capacity $C_i(R)$ of the restored node as shown in Equation (3.7). Where, β_1 is the capacity parameter of the restored node, $0 < \beta_1 < 1$, the larger β_1 is, the larger the capacity of the restored node is, and the more difficult it is for the node to fail twice.

$$C_i(R) = (1 + \beta_1) \cdot C_i \tag{3.7}$$

To strengthen the network, the recovered nodes need to re-work as quickly as possible to start sharing the load pressure of the rest of the nodes, and because of the load of the recovered nodes at this moment $L_i(R) = 0$, the original strategy of distributing the load proportionally according to the initial load is not applicable to the load distribution at this time, so this paper also proposes a load redistribution model for the recovered nodes, namely

$$\Delta L_{i\to j} = L_i^1 \times \frac{C_j(R)}{\Sigma_{m\in\Gamma_i} C_m + \Sigma_{n\in\Lambda_i} C_n(R)} \tag{3.8}$$

where when a node i in the network redistributes its own load to a recovered node j, then the product of the node's load and the ratio of the capacity of the neighboring node j to the capacity of all neighboring nodes is assigned as an additional load to the recovered node j. Γ_i denotes the set of normal neighboring nodes of the node i, and Λ_i denotes the set of recovered nodes among the neighboring nodes of the node i.

In real life, when a faulty node is repaired, the most common practice to avoid secondary failures is to block the load inflow from the surrounding faulty nodes until most of the faulty nodes have been restored before the network is active again. However, this does not take into account the network timeliness and network robustness. If all the restored nodes are opened to the surrounding overloaded nodes, the possibility of secondary failure of the restored nodes will be greatly increased. Therefore, this paper defines that the restored nodes selectively open to the surrounding overloaded nodes, and only select some restored nodes with larger importance for load redistribution, because restored nodes with larger importance have larger load carrying capacity, which can reasonably accommodate the load of the nodes, and also effectively reduce the probability of the restored nodes' secondary failures.

3.3. Model Analysis. The design simulation algorithm is as follows:

Step 1: Construct a logistics network for the Northwest region.

Step 2: Initialize network parameters α,β, and λ, and determine node load L_i^0 and capacity C_i.

Step 3: Perform an initial node attack and define the attacked node as an overloaded node.

Step 4: Redistribute the load of the overloaded node to its neighboring nodes in accordance with the initial load redistribution strategy, and determine whether the load of its neighboring nodes exceeds its own capacity; if the load of the node does not exceed its own capacity, go to Step 5; if the load of the node exceeds its own capacity, go to Step 6.

Step 5: The node continues to work normally, go to step 10.

Step 6: The node is also defined as an overloaded node and the node load distribution process continues by repeating step 4 and determining whether the load of the node exceeds the recovery threshold by the optimized failure model in chapter 3.2.1, if the load of the node does not exceed the recovery threshold, then go to step 7, if the load of the node exceeds the recovery threshold, then go to step 8.

Step 7: The node starts to recover according to the recovery model in chapter 3.2.2, go to step 9.

Step 8: Then the node is determined to be completely failed, go to step 10.

Step 9: Determine the load $L_i(R)$ and capacity $C_i(R)$ of the recovered node, reopen the recovered node to the network, and determine whether the recovered node will fail again due to re-working, if the recovered node fails, go to step 4.

Step 10: Update the network.

Step 11: The cascade failure process ends, and the maximum connected subgraph size of the network and the ratio of the number of network nodes failed to the total nodes are calculated.

The flowchart is shown in Fig. 3.1:

4. Simulation Verification and Empirical Analysis. The experiments use the network maximum connectivity subgraph size G and the ratio P of the number of network nodes failed to the total nodes as the evaluation metrics to measure the robustness of the network. The network maximum connectivity subgraph size G can be used to measure the overall connectivity of the network, when the network is attacked it will be divided into a number of sub-networks that are not connected to each other, and the sub-network that contains the largest number of nodes is called the maximum connectivity subgraph. The proportion of network failed nodes P indicates the degree of failure of network nodes, and the proportion of failed nodes to all nodes is used to indicate the spread of risk. To some extent, the network maximum connected subgraph size and network failure node proportion can reflect the change of network robustness, The formula is shown below.

$$\begin{cases} G = \frac{N^{'}}{N} \\ P = \frac{N^{''}}{N} \end{cases} \tag{4.1}$$

where $N^{'}$ denotes the number of nodes in the maximum connectivity subgraph of the network after the occurrence of cascade failure, $N^{''}$ denotes the number of failed nodes in the network after the occurrence of cascade failure, and N denotes the total number of nodes in the network.

When the network is affected by node failure, the connectivity within the network will change. The larger P is, the smaller G is, indicating that there are fewer nodes connected in the logistics network and the network is less robust. The smaller P is, the larger G is, indicating that the network still maintains a larger connectivity, and the network is more robust.

4.1. Simulation Verification.

4.1.1. Robustness of the Network with Different Recovery Times. To study the impact of different start recovery times on the network, the change in network robustness is examined by adjusting the start recovery time parameter T. Without considering the secondary failure of network nodes, the real supply chain network of an enterprise is taken as an example, and the above model is simulated and analyzed by pycharm, with the capacity parameters α and β taken as 1.5 and 0.5, respectively, and the adjustable parameter λ taken as 0.3, The findings from the simulation are presented in Fig. 4.1 and Fig. 4.2.

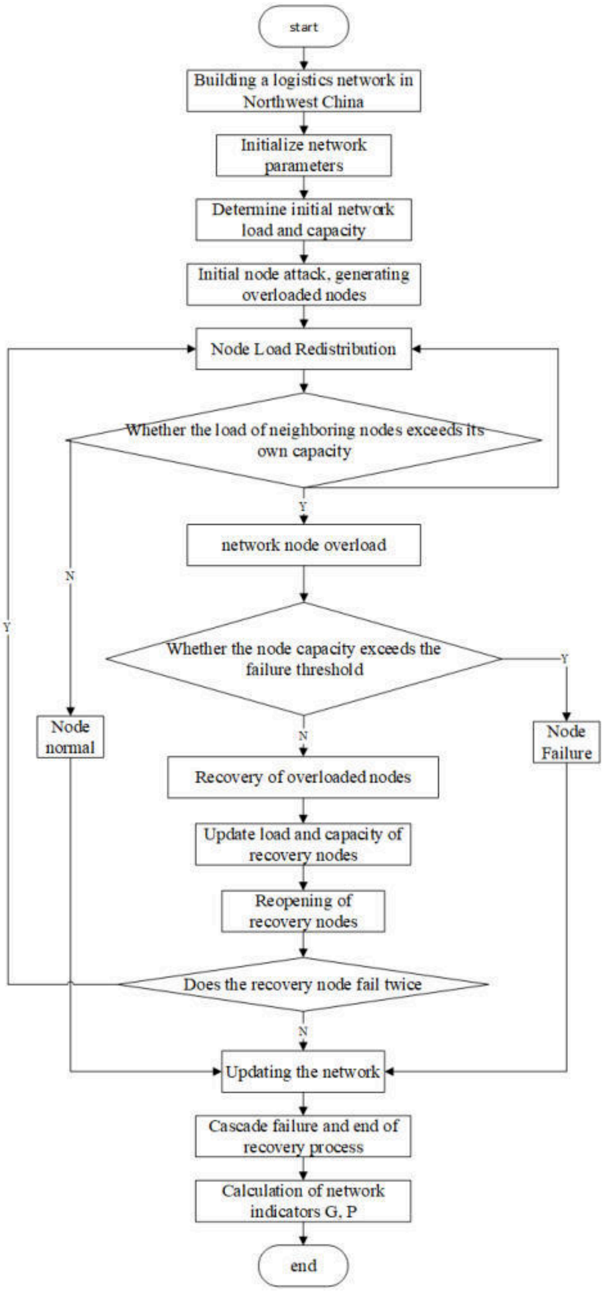

start
Building a logistics network in Northwest China
Initialize network parameters
Determine initial network load and capacity
Initial node attack, generating overloaded nodes
Node Load Redistribution
Whether the load of neighboring nodes exceeds its own capacity
Y
N
network node overload
Y
Whether the node capacity exceeds the failure threshold
Y
N
Node normal
Node Failure
Recovery of overloaded nodes
Update load and capacity of recovery nodes
Reopening of recovery nodes
Does the recovery node fail twice
N
Updating the network
Cascade failure and end of recovery process
Calculation of network indicators G, P
end

Fig. 3.1: Recovery flowchart under network cascade failure.

From Fig. 4.1 and Fig. 4.2, it can be seen that different parameters T have different impacts on the network, when T= 1, The network has the lowest proportion of failing nodes and the strongest robustness. with the increasing of T, proportion of failed nodes in the network gradually rises, and it can be known from T= 12 that when the start of the recovery time is too long, the overload nodes cannot be recovered in time, and eventually also transform into failed nodes, which causes serious harm to the operation of the network. serious harm to the operation of the network. Therefore, when the network is recovered, a smaller start recovery time should be chosen, the shorter the start recovery time is, the fewer the failed nodes will be when the network is finally

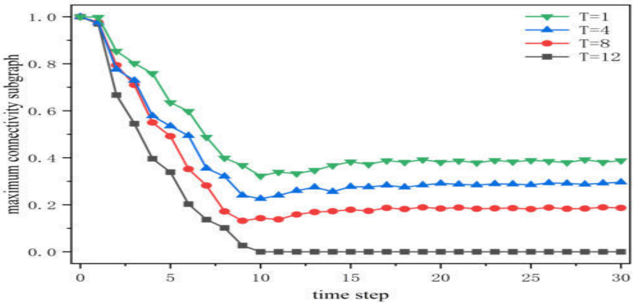


Fig. 4.1: Trend of the maximum connectivity subgraph of the network under different recovery times.

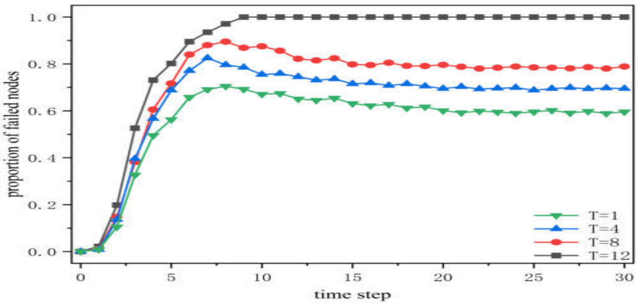


Fig. 4.2: Trend of network failed nodes under different recovery times.

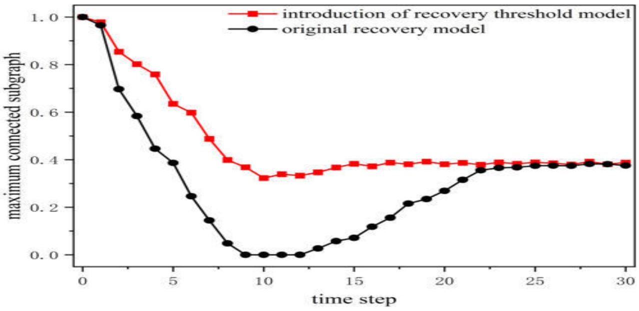


Fig. 4.3: Trend of the maximum connectivity subgraph of the network after the introduction of a recovery threshold

stabilized, and the higher the recovery ability of the network will be, and the stronger the robustness will be.

4.1.2. Introducing Robustness of Recovery Threshold Networks. On the basis of the above model, compare and analyze the change trends before and after the improvement of the recovery model under cascading fault conditions, and verify the impact of introducing node recovery threshold on the robustness of the network, the recovery time parameter is selected as $T= 1$, and the rest of the parameters are kept unchanged, and the simulation results of the original recovery model and the model with the introduction of node recovery threshold are shown in Fig. 4.3 and Fig. 4.4.

As can be seen from Fig. 4.3 and Fig. 4.4, in the original recovery model, the recovery of the failed nodes starts only after all the network nodes have failed, at which time the network has already lost the ability to

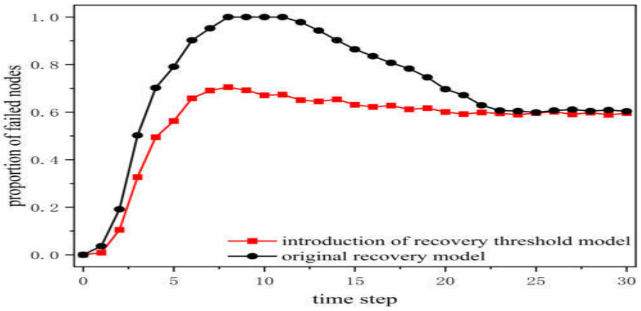


Fig. 4.4: Trend in the proportion of failed nodes in the network after the introduction of recovery thresholds

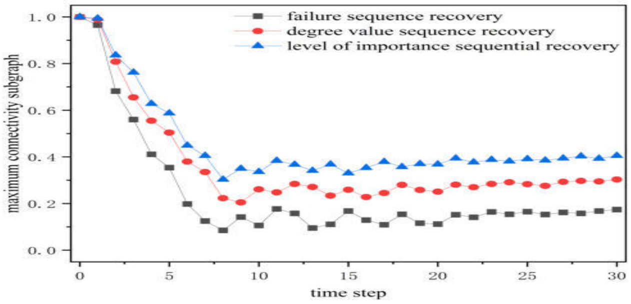


Fig. 4.5: Trend of maximum connectivity subgraph of the network under different recovery models.

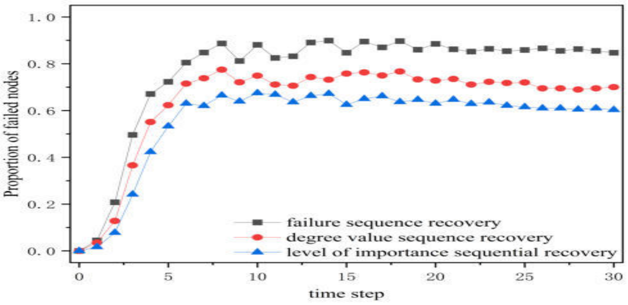


Fig. 4.6: Trend of the proportion of failed network nodes under different recovery models.

work normally, and perhaps has already caused serious impacts on human production and life. When the recovery threshold is introduced, it is equivalent to increasing the maximum capacity of the nodes, so that the nodes that should have failed are transformed into overloaded nodes instead of directly failing, and then these overloaded nodes are repaired in a timely manner. Therefore, the introduction of the recovery threshold can minimize the complete failure of the network, which can appropriately enhance the robustness of the network and avoid unnecessary losses.

4.1.3. Robustness of the Network under Different Recovery Strategies. On the basis of the above simulation experiments, considering the possibility of secondary failure of nodes, different recovery methods are compared and analyzed to verify the validity and feasibility of the model, the capacity parameter and recovery threshold parameter are kept unchanged, and the recovery time parameter is selected as $T=1$, and the simulation results of different recovery methods are shown in Fig. 4.5 and Fig. 4.6.

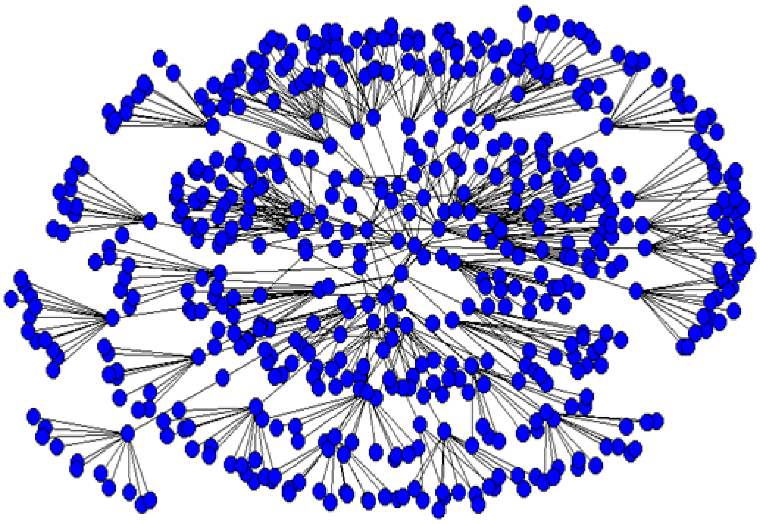


Fig. 4.7: Northwest Logistics Network.

As can be seen from Fig. 4.5 and Fig. 4.6, when the network starts to recover, the maximum connectivity subgraph curve of the network nodes and the failure ratio curve of the nodes will oscillate, fluctuating back and forth around a certain value due to the existence of the secondary failure of the nodes, and will eventually stabilize. When various conditions are the same, the effect of recovery in the order of node importance is the most obvious, and the effect of recovery in the order of network node failure is the worst, and The simulation findings show that an appropriate recovery model can improve the network's robustness.

4.2. Empirical Research.

4.2.1. Logistics Network Construction in Northwest China. This paper focuses on the logistics network in Northwest China, modeling it, cascading it to failure, and studying its robustness after recovery. Therefore, the activities of China Post Logistics in Northwest China are taken as an example to construct the postal logistics network in Northwest China. Based on the postal logistics activities of 25 municipal units, including the first-class postal district central bureau, second-class postal district central bureau, and third-class postal district central bureau established by China Post in Northwest China, the improved gravitational model establishes the logistics network between these 25 cities and then climbs a total of 25 logistics centers, 51 distribution centers and 495 business outlets within the cities to establish a total of 596 intra-city logistics activity networks. activity network, to build a total of 596 nodes of the logistics network in the northwest region, as shown in Fig. 4.7. The logistics data in this paper comes from the Postal Industry Development Statistics Bulletin of each city in 2021, and the data of logistics distribution centers and business sites are crawled from the Baidu map and the official website of China Post.

4.2.2. Characteristics of the Northwest Territories Logistics Network. On the basis of the logistics network model of Northwest China, the basic characteristic indexes of the logistics network of Northwest China are calculated by Ucinet software, and the results are shown in Table. 4.1.

The degree value, median, proximity centrality, and eigenvector centrality of the logistics network in Northwest China are calculated by Ucinet, and the sequence of the importance of the logistics network in Northwest China is obtained by hierarchical analysis based on these four indexes, and the partial ordering of these five indexes is shown in Table. 4.2.

4.2.3. Network Robustness Analysis under Recovery Policy.

(1) Effect of different recovery thresholds on network robustness. In order to study the effect of different recovery thresholds on the recoverability of the logistics network, the adjustable parameters of 0.2, 0.4, 0.6,

Table 4.1: Indicators of basic characteristics of the Northwest Territories logistics network.

Network Characteristics	calculated value
Number of nodes	596
number of connecting edges	635
average degree	2.13
network density	0.004
Average shortest path	7.03
clustering factor	0.087
Network global efficiency	0.99

Table 4.2: Comparison of various indicators of urban nodes of the logistics network in the Northwest Region.

Rankings	nodal	degree	nodal	betweenness	nodal	Proximity	nodal	eigenvector	nodal	significance
1	102	15	231	39.67	1	25.34	296	40.7	1	0.899
2	525	15	1	33.82	296	25.3	1	37.1	231	0.88
3	582	15	296	25.53	231	24.85	117	34.4	296	0.856
4	54	14	117	23.5	117	24.78	231	29.4	117	0.759
5	119	14	2	20.48	180	23.55	343	29.2	343	0.556
6	282	14	232	15.08	343	22.7	475	26	475	0.513
7	296	14	118	15.07	475	22.56	68	23.1	180	0.5
8	345	14	180	12.5	68	22.38	180	21.8	68	0.48
9	373	14	181	11.97	166	22.33	100	20.6	2	0.416
10	509	14	297	10.39	280	22.24	166	16.4	100	0.415
11	1	13	343	10.08	100	22.03	280	16.3	166	0.412
12	231	13	68	9.89	219	21.85	219	15.9	280	0.391
13	421	13	69	9.76	371	21.41	330	13.8	219	0.349
14	436	13	475	9.41	397	21.38	371	12.4	118	0.348
15	489	13	166	9.03	2	21.16	397	12.4	232	0.347
16	3	12	345	8.47	462	20.77	462	10.6	371	0.345
17	42	12	476	8.15	330	20.67	523	8.1	397	0.332
18	168	12	554	8.15	297	20.65	419	7.9	330	0.320
19	298	12	371	8.06	232	20.56	554	7.6	297	0.301
20	359	12	372	7.83	419	20.54	449	7.4	554	0.298

and 0.8 are selected for simulation verification, the capacity parameter is set as α= 1.5, β= 0.5, β_1= 0.2, and the start recovery time parameter T is taken as 1, and the simulation results are taken as the average value of 30 times. Fig. 4.8 depicts the influence of different recovery thresholds on the logistics network's maximum connectivity subgraph in Northwest China, while Fig. 4.9 depicts the impact of different healing thresholds on the proportion of failed nodes to total nodes in Northwest China.

As can be seen from Fig. 4.8, when the adjustable parameter λ= 0.2 of the recovery threshold, the size of the maximum connected subgraph of the network fluctuates around 7%, and when λ is 0.4, 0.6, and 0.8, the size of the maximum connected subgraph of the network fluctuates around 13%, 21%, and 31% and finally tends to be stabilized, which indicates that the bigger the recovery threshold, the more normal nodes there are in the maximum connected subgraph of the network. From Fig.4.9, it can be seen that when the adjustable parameter λ= 0.2 of the recovery threshold, the proportion of network failure nodes is as high as 92%, and as the parameter λ keeps increasing, the proportion of network failure nodes decreases to about 63% when λ increases to 0.8. Therefore, in the event of cascading failure faults, the higher the recovery threshold, the smaller the impact of cascading failure on the network, and the stronger the network robustness, so increasing the recovery threshold has a positive impact on improving the robustness of the logistics network in Northwest

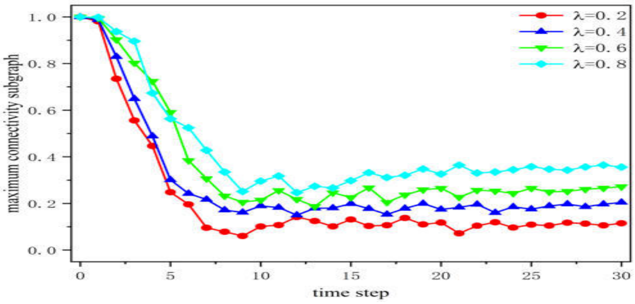


Fig. 4.8: Maximum connectivity subgraph of logistics network.

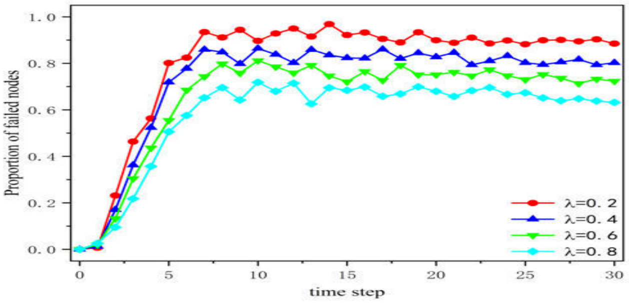


Fig. 4.9: Proportion of failed nodes in the logistics network under different recovery thresholds

China.

(2) Effect of different capacity parameters on network robustness. In order to study the influence of different capacity parameters on the recoverability of the logistics network, the adjustable parameter of recovery threshold λ is selected as 0.5, the parameter of start recovery time T is taken as 1, the capacity parameter α is taken as 1.5, β_1 is taken as 0.2, and β is taken as 0.3, 0.6, and 0.9 respectively, and simulation experiments are carried out, and the average value of 30 such results is taken. Fig. 4.10 depicts the influence of different capacity values on the logistics network's maximum connectivity subgraph in Northwest China, while Fig. 4.11 depicts the effect of various capacity parameters on the proportion of failed nodes compared to all nodes in Northwest China.

As shown in Fig. 4.10 and Fig. 4.11, when the value of λ is fixed, the different capacity parameters mainly affect the initial phase of the network, and there is less difference between the final recovery effects of the network as time passes. In both figures, $\beta = 0.3$, the network starts to fail massively at time step 1, when $\beta = 0.6$, The network's maximum connectivity subgraph begins to break severely at time step 3. and the node failure ratio of the network starts to fail massively at time step 4, whereas when $\beta = 0.9$, the network starts to fail massively at time step 5, indicating that a larger capacity parameter not only improves the robustness of the network on a small scale, but also delays the network cascade failure faults and provides some protection to the network.

(3) Recommendations for Enhancing the Robustness of Logistics Networks. Through the research and simulation of the above cascade failure and recovery strategy of the logistics network in Northwest China, It is clear that the logistical network is sturdy. in Northwest China changes with the change of the recovery threshold and capacity parameter, and the reasonable improvement of the recovery threshold and capacity parameter has a positive effect on the enhancement of the robustness of the logistics network in Northwest China. Therefore, in order to guarantee the high efficiency and stability of the logistics network in Northwest China, the following suggestions are made to suppress cascading failure faults.

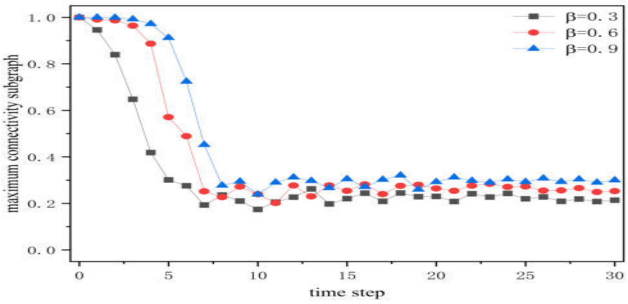


Fig. 4.10: Maximum connectivity subgraph of logistics network in Northwest China under different capacity parameters

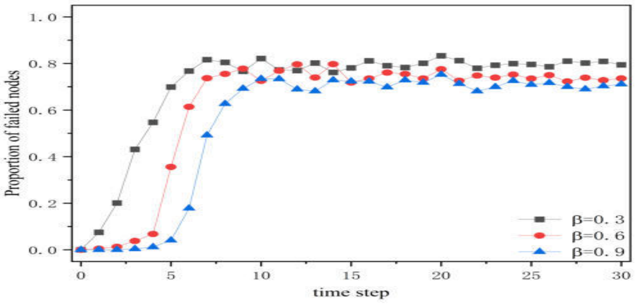


Fig. 4.11: Proportion of failed nodes in the logistics network in Northwest China under different capacity parameters

In terms of curbing cascading failures, firstly, the disaster-resistant capacity of the logistics infrastructure should be strengthened, and key nodes and facilities should be reinforced and remodeled to improve their resistance to natural disasters. Second, decentralization and redundancy strategies should be implemented to reduce the impact of a single node or path failure on the entire network by establishing backup facilities, multiple routes and other strategies. In addition, regular emergency response drills are conducted to improve practitioners' emergency response capabilities and teamwork.

In terms of recovering nodes, develop a prioritized recovery plan, set recovery priorities based on the importance of nodes, and prioritize the recovery of nodes that have the greatest impact on the entire network. At the same time, establish emergency resource reserves, including manpower, materials and technical support, so that they can be quickly put into recovery work in the event of a failure. Finally, a real-time monitoring and early warning system is established to continuously monitor the operational status of the logistics network, so that failure points can be discovered and localized in a timely manner so that recovery measures can be taken quickly.

5. Conclusions. Logistics networks need to resume normal work as fast as possible when encountering unexpected events. Therefore, it is of practical significance to take the logistics network in Northwest China as an example to conduct an empirical study to analyze its cascade failure fault and recovery behavior. In this paper, based on the original failure model, the concept of node recovery threshold is introduced to optimize the recovery time lag. Based on the original recovery model, a criticality-first recovery model is proposed, which defines the capacity of the recovered node as a function related to the original capacity, and selectively opens the recovered node to the surrounding critical nodes to reduce the risk of secondary node failure. The relationship between capacity parameters and recovery thresholds and the robustness of the logistics network in Northwest China is verified through simulation experiments, and suggestions are made to enhance the robustness of the logistics network in Northwest China. Of course, there are some limitations in this paper. In reality, the validity of the model depends on real-time accurate data, but it is often difficult to obtain such data. Monitoring the network load in real time and adjusting the recovery strategy accordingly requires an accurate control system. Limited resources may make it difficult for computational and technical requirements to be met, leading to more difficult implementations. To solve the above problems, we can improve the technology of data acquisition to enhance the real-time and accuracy of data, or we can cooperate with other companies to share data resources to increase the availability of data and reduce the cost of acquiring it, and develop modular control systems that allow for flexible adjustments and upgrades to adapt to different environments and demands, and in terms of cost, maximize the use of the limited resources. The technologies and limitations faced in reality can be overcome to a certain extent by the above mentioned ways, in order to ensure that the entire logistics network in the Northwest Territories can function properly and guarantee the smooth flow of goods in case of risks.

Funding. Natural Science Foundation of Gansu Province (Project No. 23JRRA898)

REFERENCES

[1] Zhang W D, Yin Y, Gu W J et al., *Association network modeling of multidimensional smart shop floor data in complex spatio-temporal domain*, Control and Decision Making, 2024, 1-9.

[2] Lai Q, Ma X, Zhang HH, et al., *Robustness analysis of aviation network structure*, Journal of Huazhong University of Science and Technology (Natural Science Edition), 2023, 1-6.

[3] Li R, Chen X, *Reverse Logistics Network Design under Disruption Risk for Third-Party Logistics Providers*, Sustainability, 2022, 14(22):14-31.

[4] Yu J, Xiao B, Cui Y Z, *Recovery method of group-dependent combat network under overload situation*, Journal of Harbin Institute of Technology, 2024, 1-11.

[5] Cheng G , Hongyi P , Mengchao W , et al., *Identifying priority areas for ecological conservation and restoration based on circuit theory and dynamic weighted complex network: A case study of the Sichuan Basin*, Ecological Indicators, 2023, 155:45-57.

[6] Yang Q, Zhang Y N, Zhou Y Q et al., *A review of complex network theory and its application in public transportation resilience*, China Highway Journal, 2022, 35(04):215-229.

[7] Qin L ,Hongkai C ,Yuhan L , et al., *Network spreading among areas: A dynamical complex network modeling approach*, Chaos (Woodbury, N.Y.), 2022, 32(10):103102-103102.

[8] Yang B, Qian X D, *Research on the evolution of supply chain enterprise cooperation based on complex network model*, Complex Systems and Complexity Science, 2018, 15(03):1-10.

[9] Chen D Q, Sun D Z, Yin Y Q, Dhamotharan L, Kumar A, Guo Y H, *The resilience of logistics network against node failures*, International Journal of Production Economics, 2022, 244-255.

[10] Wang Y D, Shi Q, Song S, Hu Q W, *Multi-objective optimization model and solution algorithm for dynamic logistics network*, Computer Integrated Manufacturing Systems, 2020, 26(04):1142-1150.

[11] Zhang D Z, He R Z, Zhu W, Zhang Z, *Research on optimization of multi-mode express logistics service network based on low carbon economy perspective*, Journal of Railway Science and Engineering, 2018, 15(06):1601-1608.

[12] Qian X D, Xuan Z, *Research on the diffusion of risk factors in logistics networks under the background of complex networks*, Computer Engineering and Application, 2022, 58(13):303-314.

[13] Wang S H, Yang Y, Sun L Y, Li X N, Li Y X, Guo K H, *Controllability Robustness Against Cascading Failure for Complex Logistic Network Based on Dynamic Cascading Failure Model*, IEEE Access, 2020, 8, 34-47.

[14] Zhang Z, Liu D Y, Zhang J et al., *Robustness study of complex networks based on dynamic control of node redundancy capacity*, Journal of Electronics and Information, 2021, 43(05):1349-1356.

[15] German R M, Daniel M, John E W, *A new method for the measurement of robustness in reverse logistics supply chains based on entropy and nodal importance*, Computers & Industrial Engineering, 2023, 183, 121-135.

[16] Haihua Y, Shi A, *Robustness evaluation for multi-subnet composited complex network of urban public transport*, Alexandria Engineering Journal, 2021, 60(2):64-78.

[17] Yu R B, Jiang Y, Yan Y W et al., *Robustness study of dependent networks considering dependent edge loading*, Journal of University of Electronic Science and Technology, 2022, 51(05):774-785.

[18] Zhang L, Du Y, *Resilience of space information network based on combination of complex networks and hypergraphs*, Computer Communications,2022,195,124-136.

[19] Yang Y, Peng C, Yang Y J et al., *Research on supply chain recovery strategy under disruption risk*, Journal of System Simulation, 2021, 33(12):2771-2781.

[20] Wu J ,Chen Z ,Zhang Y , et al., *Sequential Recovery of Complex Networks Suffering From Cascading Failure Blackouts*, IEEE Transactions on Network Science and Engineering, 2020, PP(99).

[21] Ju Y N, Li Z P, Chen Y F, et al., *Study on node importance and fault recovery of regional rail transit systems*, Chinese Journal of Safety Science, 2021, 31(02):112-119.

[22] Tang L, Jiao P, Li J K, et al., *Research on cascading failure mechanism and robustness of complex networks with recovery strategies*, Control and Decision Making, 2018, 33(10):1841-1850.

[23] Huang Y, Liu M R, Wei J G, et al., *Study on the recovery strategy of urban subway network based on resilience curve*, Disaster Science, 2021, 36(01):32-36.

[24] Xiuwen F, Xiaojie X, Wenfeng L, *Cascading failure resilience analysis and recovery of automotive manufacturing supply chain networks considering enterprise roles*, Physica A: Statistical Mechanics and its Applications, 2024, 634129478-634129489.

[25] Yoshihisa S, Shinsuke M, *Designing a resilient international reverse logistics network for material cycles: A Japanese Case study*, Resources, Conservation Recycling, 2021, 170:25-36.

[26] Motter A E, Lai Y C, *Cascade-based attacks on complex networks*, Physical Review E, 2002, 66(6):76-89.

[27] Dong-Hee Kim, Adilson E Motter, *Resource allocation pattern in infrastructure networks*, Journal of Physics A: Mathematical and Theoretical, 2008, 41(22):97-109.

[28] Wang S C, Qian X D, *Optimization of cascading failure model and network robustness analysis based on complex networks-an example of postal logistics network in Northwest China*, Control Engineering, 2024, 1-11.

[29] Di Muro M A, La Rocca C E, Stanley H E, et al., *Corrigendum: Recovery of Interdependent Networks*, Scientific reports, 2017, 7(1-4):46586.

[30] Böttcher, Lucas, et al., *Failure and recovery in dynamical networks*, Scientific reports, 2017, 7(1-4):41729.

[31] Hu B, Li F, *Repair strategies of scale-free networks under multifold attack strategies*, Systems Engineering and Electronics, 2010, 32(1):86-89.

[32] Jiarong X, Youyou Y, Zhengping F, et al., *Eradicating abrupt collapse on single network with dependency groups*, Chaos (Woodbury, N.Y.), 2019, 29(8):083111.

[33] Xin Y, Yanqing H, Eugene H S, et al., *Eradicating catastrophic collapse in interdependent networks via reinforced nodes*, Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13):3311-3315.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Mar 18, 2024
Accepted: May 20, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 151–159, DOI 10.12694/scpe.v26i1.3722

PEPTIDE SEQUENCE TAG EXTRACTION BY GRAPH CONVOLUTION NEURAL NETWORKS

XINYE BIAN[*] DONGMEI XIE[†] DI ZHANG[‡] XIAOYU XIE[§] YUYUE FENG[¶] PIYU ZHOU[‖] CHANGJIU HE[**]MINGMING LYU[††]AND HAIPENG WANG[‡‡]

Abstract. The peptide sequence tag extraction method plays a vital role in tandem mass spectrometry-based protein identification engines. This approach faces two significant challenges in practical applications: first, the issue of fixed tag lengths, where shorter tags lack sufficient specificity, leading to an excessive recall of non-target peptide sequences, and longer tags experience a reduction in precision as tag length increases, potentially failing to recall target peptide sequences; second, the sensitivity and precision of tag extraction remain relatively low. To address these issues, a variable-length peptide sequence tag extraction algorithm, TagEx, based on graph convolutional networks, is proposed. This method begins by training a de novo peptide sequencing scoring model utilizing graph convolutional networks. It then constructs a spectral peak connection graph from the mass spectrum, employing a depth-first search strategy to extract variable-length peptide sequence tags, with the trained graph convolutional network model scoring amino acid connections during the extraction process.Finally, tags are filtered based on length and scoring to obtain the final candidate peptide sequence tag set. To evaluate TagEx's performance, it was benchmarked against three representative tag extraction software tools: InsPect, PepNovo+, and DirecTag. The experimental results demonstrate that TagEx exhibits superior sensitivity, coverage, and precision, with improvements of 0.62-2.32, 3.22-11.14, and 3.29-8.31 percentage points, respectively, when retaining the top 100 tags.

Key words: proteomics,peptide sequence tag, graph convolutional neural network, de novo sequencing, tandem mass spectrometry

1. Introduction. With the continuous advancement of scientific and technological progress, the precision of mass spectrometry instruments has steadily improved, making tandem mass spectrometry analysis an indispensable key technology for protein identification. Tandem mass spectrometry analysis can generally be categorized into three primary methods, including database-based searching, de novo sequencing, and tag-based database searching method[1]. Among these, the protein database search method is the most commonly used for protein identification, and widely used database search software includes MS-GF+[2], pFind-Alioth[3], Comet[4]and others. The method mainly refers to existing mass spectrometry data and peptide sequences in protein databases; therefore, a major drawback of the database search method is that it is unable to explore proteins in unknown areas. Another common method for protein identification is de novo sequencing, which no longer depends on protein databases but infers peptide sequences directly from spectral information. Common de novo sequencing tools include SMSNet[5] , PointNovo[6], Casanovo[7], denovoGCN[8], and more. During the de novo sequencing process, the entire sequence corresponding to the spectrum needs to be predicted, and even if there are slight differences from the actual peptide sequence, it is considered an incorrect sequence. However, these incorrect sequences may contain some correct fragment sequences that still have value in peptide spectrum matching. Therefore, the sequence tag method is an approach that combines the strengths of both database searching and de novo sequencing methods.

[*] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[†] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[‡] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[§] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[¶] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[‖] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[**] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[††] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[‡‡] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China (Corresponding author, hpwang@sdut.edu.cn)

The peptide sequence tagging method is a technique utilized in proteomics mass spectrometry analysis for the identification of peptide sequences. This method initially infers partial fragments of peptides, namely peptide sequence tags, from tandem mass spectrometry data. Subsequently, it searches a protein database for candidate peptide sequences containing these sequence tags to achieve the final peptide spectrum matching results[9]. Within the field of proteomics research, this approach holds significant value for the rapid identification of peptide sequences, as well as the discovery and localization of unknown modifications.

The concept of peptide sequence tagging can be traced back to 1994, initially proposed by Mann et al[10]. They defined sequence tags as segments of consecutive fragment ions within tandem mass spectra and used these sequence tags as keywords to retrieve corresponding peptide sequences from protein sequence databases. In 2003, Tabb et al. published the first standalone sequence tag extraction software, GutenTag, which utilized statistical methods to generate a model of spectral peak relative intensity, thereby facilitating large-scale extraction of sequence tags. In the same year, Sunyaev et al. introduced MultiTag, employing a fault-tolerant sequence tagging approach to recall homologous proteins and perform statistical evaluations. In 2005, Tanner et al[11]. developed the InsPect protein search engine, primarily utilizing dynamic programming for peptide sequence tag extraction and leveraging these tags to filter retrieved proteins. Concurrently, Frank et al. published the de novo sequencing tool, PepNovo, adopting probabilistic network modeling combined with graph theory and dynamic programming to infer amino acid sequences from mass spectral data. In 2008[12], Tabb et al. introduced a new sequence tag extraction tool, DirecTag, which generated sequence tags through recursive enumeration and scored tags based on intensity rank, error variance, and ion complementarity, demonstrating improved performance in sequence tag extraction compared to GutenTag. In 2009, Frank et al[13]. enhanced PepNovo by introducing a new scoring function based on adaptive boosting, resulting in PepNovo+. However, since then, standalone sequence tag extraction tools have become relatively scarce, with most tools being integrated within protein search engines without offering an independent sequence tag output interface, such as MODa[14], Open-pFind[15], and MODplus[16], among others.

Currently, the majority of sequence tag extraction tools are configured with fixed tag extraction lengths. To reduce the difficulty of extracting correct tags and to ensure the recall rate of target peptide sequences, tag lengths are typically constrained to a range of three to five amino acids. However, setting a fixed tag length presents certain disadvantages. When the sequence tag length is set too short, the specificity of the tags may be insufficient, leading to the recall of numerous non-target peptide sequences; conversely, longer tags may decrease in precision as their length increases, potentially failing to recall target peptide sequences. At the same time, due to the outdated algorithms used for scoring tags, the sensitivity and precision of most current tag extraction tools remain low, significantly impacting the effectiveness of tag extraction methods in proteomics data analysis.

This article introduces a tag extraction method named TagEx, based on Graph Convolutional Neural Networks (GCN). This method initially involves training a peptide sequencing scoring model utilizing graph convolutional networks. Subsequently, a spectral peak connection graph is constructed based on the mass spectrum, employing a depth-first traversal strategy to extract variable-length peptide sequence tags. During the tag extraction process, the trained graph convolutional network model is utilized to score amino acids on connecting edges. Finally, tags are filtered based on their length and scores to obtain the final candidate peptide sequence tag set. By integrating a variable-length tag extraction algorithm with the GCN model, TagEx exhibits outstanding performance in terms of tag sensitivity, coverage, and precision.

2. Materials and methods.

2.1. Dataset description. TagEx employs high-precision data comprising 1528127 spectra from nine species, including V.mungo, M.musculus, M.mazei, C.endoloripes, S.lycopersicum, S.cerevisiae, A.mellifera, H.sapiens, Bacillus, as the dataset. The first eight species are utilized to train the de novo sequencing model based on graph convolutional neural networks. The original data for these species are sourced from the PRIDE database, specifically from PXD005025[17], PXD004948[18], PXD004325[19], PXD004536[20], PXD004947[21], PXD003868[22], PXD004467[23] and PXD004424[24]. The data from another species not included in the training set, PXD004565[25], is used as the test set. All datasets mentioned are uniformly exported as MGF files using the pParse+ software, with peptide sequence identification performed using PEAKS software to obtain peptide spectrum matching information. The data is filtered using a False Discovery Rate (FDR)[26],

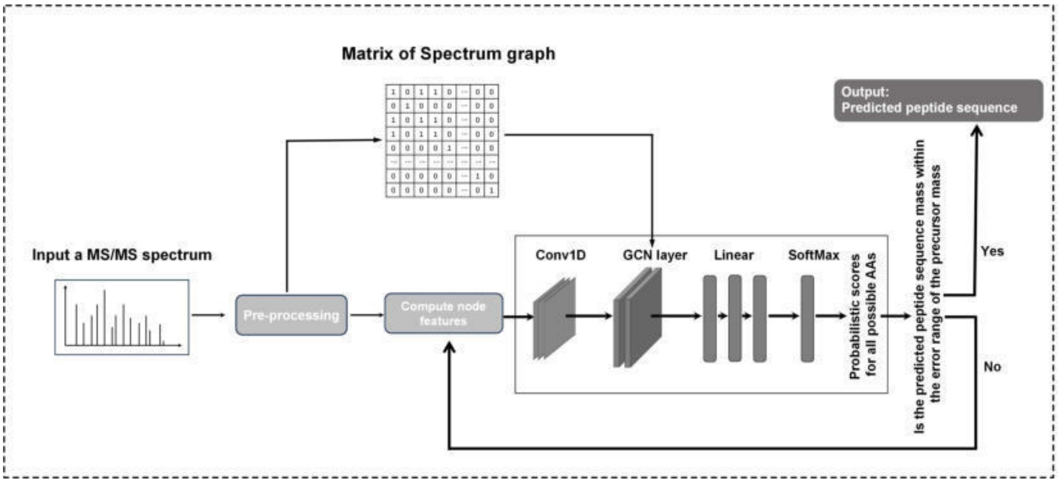

Matrix of Spectrum graph
Output:
Predicted peptide sequence
Input a MS/MS spectrum
Pre-processing
Compute node features
Conv1D
GCN layer
Linear
SoftMax
Probabilistic scores for all possible AAs
Is the predicted peptide sequence mass within the error range of the precursor mass
Yes
No

Fig. 2.1: De novo sequencing model training process.

the threshold of 0.01.

2.2. De novo sequencing model training based on graph convolutional networks. The Graph Convolutional Neural Network (GCN) was proposed by Kipf and Welling[27] in 2017 and swiftly became a seminal model within the domain of graph neural networks. Its foundational principle lies in conducting convolution operations on nodes and their adjacent nodes, thereby facilitating the extraction from local to global features. This feature enables GCNs to not only preserve graph structural information but also significantly enhance performance in various graph data tasks. The excellence of GCNs in tasks such as node classification and graph classification can be attributed to their convolution operations.

For the convenience of model performance testing, the training set is divided into an 8:2 ratio, with 80% allocated for training purposes and the remaining 20% serving as the validation set. The specific training process follows that of denovoGCN, as proposed by our research group. The initial step involves preprocessing the spectra, including the addition of four virtual peaks, converting the m/z ratio to neutral mass, and preserving spectral peaks among other operations. Subsequently, a mass spectrum connection matrix is constructed, and spectral peak features are developed based on the m/z ratio and intensity information of the currently predicted peptide sequence. Finally, the features and the mass spectrum connection matrix are input into the model to obtain probability estimates for all possible identities of the next amino acid, with the amino acid having the highest probability added to the predicted sequence. The sequence is output as the final de novo sequencing result once the mass of the predicted peptide sequence falls within the error range of the parent ion mass. The model training process is depicted in Figure 2.1.

2.3. Tag Extraction. Upon successful training of the model, it is utilized for tag extraction through the following specific process: 1) Spectrum preprocessing: conversion of spectral peak intensities to relative intensities; addition of four virtual peaks to the spectrum with m/z ratios equivalent to the mass of a proton, the mass of a singly charged water molecule, the mass of a singly charged parent ion, and the mass of a dehydrated singly charged parent ion, each with an intensity of 1; retention of the top 200 peaks with the highest intensities. 2) Construction of the spectral peak connection graph: starting from the first spectral peak in the preprocessed spectrum, all peaks are traversed. The mass difference between the current peak and other peaks with higher m/z ratios is calculated, and if the mass difference matches the mass of an amino acid, these two peaks are connected. 3) Amino acid edge scoring: during the traversal of spectral peaks, the mass of the traversed peak is input into the model to obtain scores for amino acids on connecting edges starting from the current peak. This continues until traversal is complete, resulting in a spectral peak connection graph with scored connecting amino acids. 4) Tag extraction: initially, a buffer is established to store the extracted amino

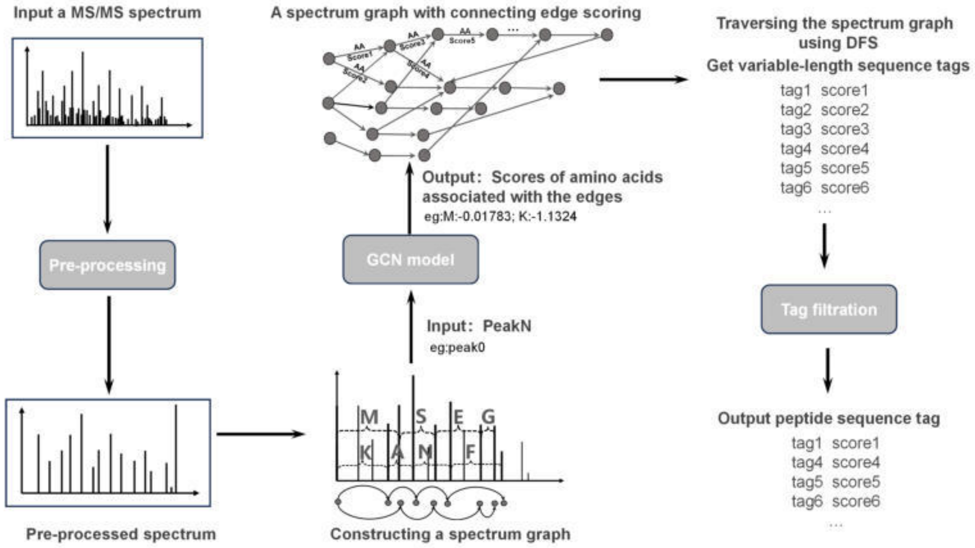


Fig. 2.2: Peptide sequence tag extraction flowchart.

acid sequence information. Then, all nodes with an indegree of zero are traversed, performing a depth-first traversal of the path from that node. During the depth-first traversal, the average of the scores of the amino acids on the traversed edges and those in the buffer is calculated to determine if this average is within the set scoring threshold t. If higher than the threshold, the amino acid and its score corresponding to the current edge are added to the buffer. If lower than the threshold, the length of the amino acid sequence in the buffer is assessed; if greater than 3, it is saved as a peptide sequence tag and the average of the amino acid scores is taken as the tag's score, then the tag and its score are saved before clearing the buffer; otherwise, the buffer is cleared directly. When traversal reaches a node with an outdegree of 0, the amino acid sequence in the buffer is saved as a tag and backtracking occurs. 5) Tag filtering: tags of varying lengths are filtered using different empirical scoring thresholds based on tag length, and the filtered results constitute the final candidate peptide sequence tag set. The peptide sequence tag extraction process is illustrated in Figure 2.2.

2.4. The performance evaluation indicators. TagEx utilizes three performance metrics to evaluate the tag extraction algorithm, including sensitivity[28], coverage[29], and precision. Initially, sequence tags extracted by various software are ranked based on their scoring, with the top-scored tags being selected sequentially. Sensitivity is defined as the proportion of spectra containing correct tags within the top n tags to the total number of spectra. The formula for calculating tag sensitivity can be represented by Equation (2.1), where s denotes the number of spectra from which correct tags can be extracted, and S represents the total number of spectra tested.

$$sensitivity = \frac{s}{S} \tag{2.1}$$

For the evaluation metric of coverage, consider a specific spectrum and its corresponding peptide sequence ”PEPTIDESEQ” as an example. Suppose that the first tag extracted by the tag extraction algorithm is ”PEPTI” and the second tag is ”EPTID”. In this scenario, the coverage for the top 2 tags of this spectrum is calculated as the number of covered amino acids, 6, divided by the total length of the peptide sequence, 10, resulting in a coverage value of $\frac{6}{10} = 0.6$ for the top 2 tags of this spectrum. Subsequently, by calculating the coverage for all spectra in the dataset and then averaging these values, the overall dataset coverage is determined. Furthermore, in the assessment of coverage, the distribution of coverage by extracted tags when selecting the top 100 tags in each spectrum is also considered. This implies that not only the coverage of individual tags is calculated, but a comprehensive evaluation is also performed on the amino acids covered by all tags within each spectrum,

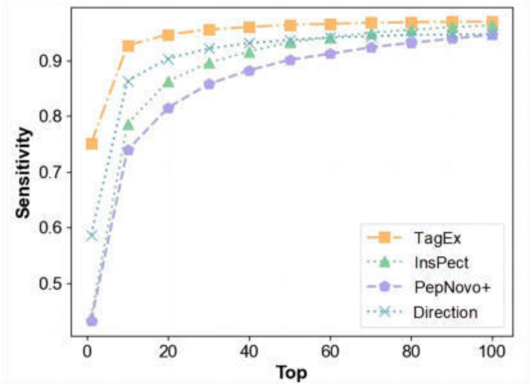


Fig. 3.1: Comparison of tag extraction sensitivity.

leading to a more holistic analysis of coverage. The formula for tag coverage can be denoted by Equation (2.2), where c represents the total coverage of all spectra in the dataset, and C denotes the number of spectra.

$$coverage = \frac{c}{C} \tag{2.2}$$

The precision of a spectrum's tags is measured by the proportion of correct tags within the top n selection. The precision across the test dataset is calculated as the mean precision for all spectra from which tags can be extracted. The formula for tag precision can be denoted by Equation (2.3), where p is the sum of precision for the spectra, and P is the count of spectra from which tags can be extracted.

$$precision = \frac{p}{P} \tag{2.3}$$

3. Experiment and Result Analysis.

3.1. Comparison Software Setup. To demonstrate the performance of TagEx in tag extraction, this article conducts a comparative analysis of TagEx with three other tag extraction tools—InsPecT, PepNovo+, and DirecTag—on the PXD004565 dataset. Since DirecTag requires input files in mzML format, it is necessary to first utilize the msConvert software to convert MGF files exported by the pParse+ software into the requisite mzML format. To ensure fairness in comparison, all tag extraction tools were configured to include only the fixed modification of Carbamidomethyl on amino acid C, and the amino acid mass error value was uniformly set to 0.02 Da. During the comparison process, the tag extraction lengths for the other three tools were set to 3 to enable optimal performance in tag extraction.

3.2. Sensitivity Comparison Experiment. On the PXD004565 dataset, the sensitivity performance of each tool is depicted in Figure 3.1. When selecting the top 1 tag from all software, TagEx achieves a sensitivity of 75.01%, which is respectively 16.38, 31.8, and 31.76 percentage points higher than DirecTag, InsPect, and PepNovo+. When selecting the top 100 tags, TagEx's sensitivity reaches 96.88%, which is respectively 2.23, 0.62, and 2.32 percentage points higher than DirecTag, InsPect, and PepNovo+. Throughout the entire process where sensitivity rises and eventually stabilizes, TagEx consistently maintains a leading advantage compared to the other tag extraction tools.

3.3. Coverage Comparison Experiment. On the PXD004565 dataset, the coverage comparison curves of various tag extraction tools are illustrated in Figure 3.2. Since TagEx employs a variable-length tag extraction method, unlike other tools that utilize fixed-length tags, it is not feasible to fairly compare coverage at the top 1 tag; hence, the coverage for the top 1 tag is not recorded. When selecting the top 20 tags from all software, TagEx's coverage reaches 57.21%, which is respectively 8.58, 20.85, and 17.92 percentage points higher than DirecTag, InsPect, and PepNovo+. Upon selecting the top 100 tags from all software, TagEx's

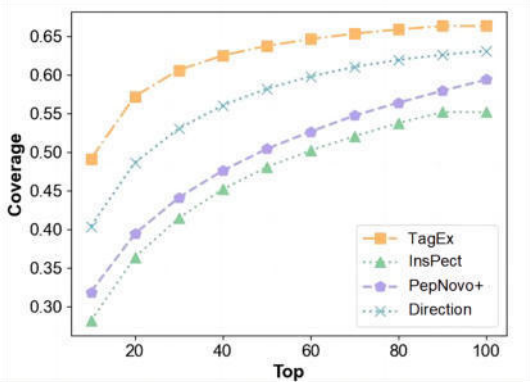


Fig. 3.2: Comparison of tag extraction coverage.

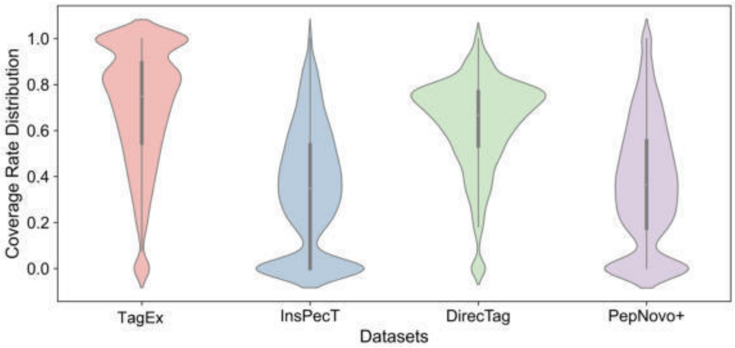


Fig. 3.3: Distribution of per-spectrum coverage in the Top100 tags by software.

coverage attains 66.33%, which is respectively 3.22, 11.14, and 6.94 percentage points higher than DirecTag, InsPect, and PepNovo+. Consequently, TagEx demonstrates superior performance in terms of coverage.

In the evaluation of tag coverage, the average value across all spectra is utilized, which may disadvantage tag extraction tools that exhibit low coverage in only a small subset of spectra. Therefore, after selecting the top 100 tags for each tag extraction tool, the distribution of coverage for each spectrum was analyzed, as shown in Figure 3.3. The coverage for each spectrum by TagEx is predominantly distributed around 100% and 80%, whereas the coverage for each spectrum by InsPect and PepNovo+ is mostly distributed within the 0-20% range, and the coverage for each spectrum by DirecTag is largely distributed within the 60-80% range. Hence, from the perspective of coverage per spectrum, TagEx also demonstrates an advantage.

3.4. Precision Comparison Experiment. Within the PXD004565 dataset, the precision of tags extracted by various tag extraction tools is depicted in Figure 3.4. When selecting the top 20 tags from each tag extraction tool, TagEx achieves a precision of 41.05%, which is respectively 10.91, 21.108, and 25.752 percentage points higher than DirecTag, InsPect, and PepNovo+. Upon selecting the top 100 tags from each tag extraction tool, the tag precision of TagEx is 14.56%, which is respectively 6.735, 3.287, and 8.313 percentage points higher than DirecTag, InsPect, and PepNovo+. Thus, TagEx demonstrates exceptional performance in terms of precision.

The precision of extracted tags decreases as the length of the tag increases. To further elucidate TagEx's performance in terms of precision, a comparison was made between the precision of sequence tags of lengths 3, 4, and 5 extracted by TagEx and those extracted by other tag extraction tools of corresponding lengths, with results depicted in Figure 3.5. Since DirecTag is only capable of extracting tags of lengths 3 and 4, its data

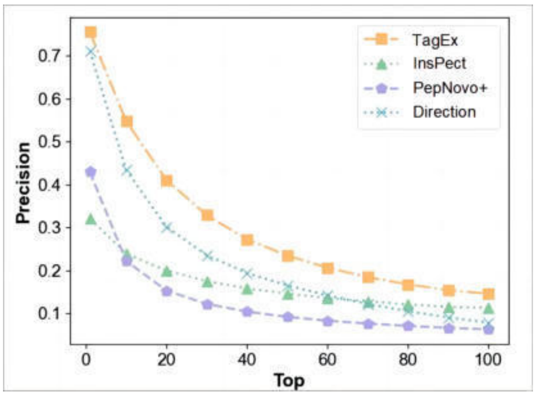


Fig. 3.4: Comparison of tag extraction precision.

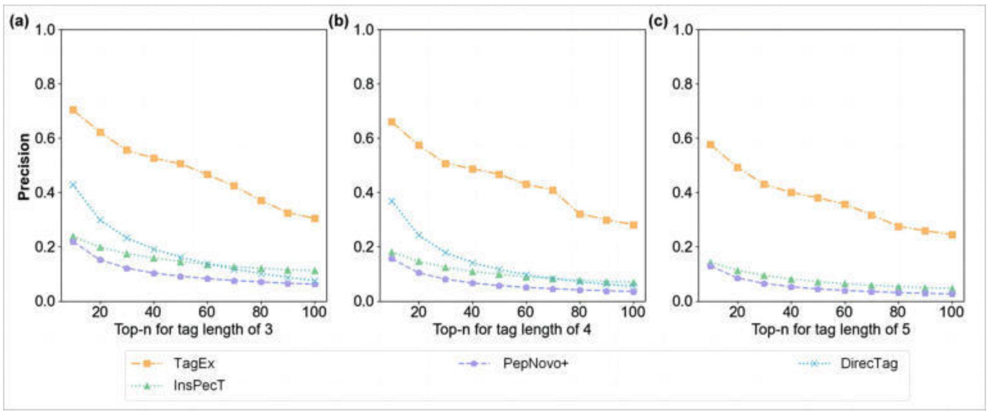


Fig. 3.5: (a) Precision comparison of tags of length 3. (b) Precision comparison of tags of length 4. (c) Precision comparison of tags of length 5.

is not included in the comparison for length 5 tags. Figure 3.5(a) displays the comparison results for length 3, where TagEx achieves a precision of 30.39% for tags of length 3 when selecting the top 100 tags, which is 19.1 to 24.14 percentage points higher than other software. Figure 3.5(b) shows the comparison results for length 4, with TagEx achieving a precision of 28.14% for tags of length 4, which is 17.06 to 21.99 percentage points higher than other software. Figure 3.5(c) presents the comparison results for length 5, where TagEx achieves a precision of 24.17% for tags of length 5, which is respectively 23.19 and 25.74 percentage points higher than InsPect and PepNovo+. Therefore, even when comparing tag precision segmented by tag length, TagEx still exhibits the best performance.

TagEx demonstrates outstanding precision performance, primarily due to our use of a de novo sequencing model based on graph convolutional networks as the scoring model for tags. This method more effectively ensures that the correct tags extracted by TagEx receive higher scores, thereby enhancing the overall precision of tag identification. The advantage of TagEx becomes more apparent when comparing precision based on tag length because the difficulty of determining tags increases with length, inevitably leading to a reduction in precision. Unlike other tag extraction tools that use fixed lengths of 3 to 5 amino acids, TagEx extracts tags of variable lengths, all of which are at least three amino acids long. Therefore, TagEx is at a comparative disadvantage when evaluated against tools that use a fixed tag length. However, when the tags extracted by TagEx are analyzed by length and compared for precision against other tools, their superiority becomes more evident.

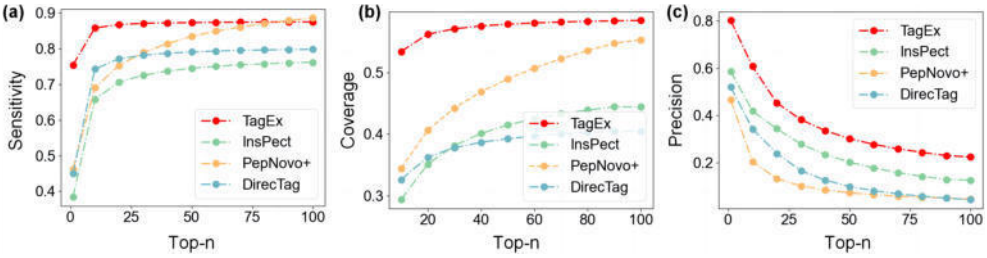


Fig. 3.6: (a) Comparison of sensitivity in the PXD009449 dataset. (b) Comparison of coverage in the PXD009449 dataset. (c) Comparison of precision in the PXD009449 dataset.

3.5. Software Performance Comparison in the PXD009449 Dataset. In the PXD009449 dataset, this study conducted a comparative analysis of the performance of various tag extraction software. The primary purpose of this analysis was to evaluate the capability of TagEx in extracting tags within complex modification environments and to ensure the algorithm's generalizability across different datasets, thus preventing its performance from being limited to specific datasets.

In terms of sensitivity, TagEx demonstrated significant superiority, with improvements of 12.19 to 38.86 percentage points and 8.57 to 32.25 percentage points over InsPect and DirecTag, respectively. Additionally, when selecting the Top80 tags, GCNTag showed an increase of 0.29 to 29.98 percentage points compared to PepNovo+. However, in the selection of Top90 to Top100 tags, PepNovo outperformed GCNTag, exhibiting higher sensitivity by 0.51 to 1.21 percentage points as the Figure 3.6(a). In terms of tag coverage, TagEx also performed better than InsPect and DirecTag, with coverage increases of 9.95 to 21.55 percentage points, 2.79 to 18.43 percentage points, and 13.91 to 19.01 percentage points as the Figure 3.6(b). Regarding precision, the improvements in TagEx were 22.23 to 35.15 percentage points, 27.14 to 45.11 percentage points, and 26.75 to 41.3 percentage points compared to InsPect and DirecTag as the Figure 3.6(c).

Therefore, based on the results of the comparative experiments, it is evident that GCNTag still exhibits a significant advantage in overall tag extraction performance on the PXD009449 dataset compared to other tag extraction tools.

4. Conclusions. This article introduces a variable-length tag extraction method to address the issues posed by fixed-length sequence tags and employs a graph neural network model to fit peptide spectrum matching patterns as a scoring tool to enhance the sensitivity and precision of the extracted tags. To evaluate TagEx's performance, it was benchmarked against three representative tag extraction software tools: InsPect, PepNovo+, and DirecTag. The experimental results demonstrate that TagEx exhibits superior sensitivity, coverage, and precision, with improvements of 0.62-2.32, 3.22-11.14, and 3.29-8.31 percentage points, respectively, when retaining the top 100 tags; the advantages are even more pronounced when only the top-ranked tag is retained, with sensitivity and precision increasing by 16.38-31.76 and 4.387-32.597 percentage points, respectively.

In summary, in comparison with three representative tag extraction software, the tags mentioned by TagEx exhibit advantages in sensitivity, coverage, and precision. Moreover, this variable-length peptide sequence tag extraction method provides a more flexible and accurate algorithmic foundation for subsequent rapid peptide sequence identification, discovery, and localization of unknown modifications.

REFERENCES

[1] J.Yuming,R.Devasahayam,S.Dina,N.Benjaminand,V.Norbert,M.Amanda,Peters-Clarke,T.M,E.Susan,K.Simion, *pFind–Alioth:A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data*, Journal of Proteomics,125(2015),pp. 89-97.

[2] K.Sangtae and P.PavelA, *MS-GF+ makes progress towards a universal database search tool for proteomics*, Nature Communications,2023.

[3] C.Hao,H.Kun,Y.Bing,C.Zhen,S.Rui-Xiang,F.Sheng-Bo,Z.Kun,L.Chao,Y.Zuo-Fei,W.Quan-Hui, *Comprehensive Overview of Bottom-Up Proteomics using Mass Spectrometry*, ArXiv,125(2015),pp. 89-97.

[4] B.Navratan,B.Elena,C.Enrique,L.AnaVictoria,M.Spiros,T.Marco,E.Iakes,J.Manuel, M.Ricardo,L.Ana, *Comprehensive quantification of the modified proteome reveals oxidative heart damage in mitochondrial heteroplasmy*, Cell Reports,23(2018),pp. 3685-3697.

[5] W.Jinwei,Z.Junjie,Y.Qilin,L.Xiangyang,Z.Yuhui,S.Yun-Qing,J.Sunil, *SmsNet: A New Deep Convolutional Neural Network Model for Adversarial Example Detection*, IEEE Transactions on Multimedia,24(2021),pp. 230-244.

[6] Q.Rui,T.Ngoc-Hieu,X.Lei,C.Xin,L.Ming,S.Baozhen,G.Ali, *Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices*, Nature Machine Intelligence,3(2021),pp. 420-425.

[7] Y.Melih,F.William,B.Wout,O.Sewoong,N.William, *De novo mass spectrometry peptide sequencing with a transformer model*, International Conference on Machine Learning,(2022),pp. 25514-25522.

[8] W.Ruitao,Z.Xiang,W.Runtao,W.Haipeng, *Denovo-GCN: De Novo Peptide Sequencing by Graph Convolutional Neural Networks*, Applied Sciences,13(2023),p. 4604.

[9] V.Kira, *Validation of de novo peptide sequences with bottom-up tag convolution*, Proteomes,10(2021),p. 1.

[10] M.Matthias,W.Matthias, *Error-tolerant identification of peptides in sequence databases by peptide sequence tags*, Analytical Chemistry,24(2021),pp. 4390–4399.

[11] T.Stephen,S.Hongjun,F.Ari,W.LingChi,Z.Ebrahim,M. Marc,P.Pavel,B.Vineet, *InsPecT: identification of posttranslationally modified peptides from tandem mass spectra*, Analytical Chemistry,77(2005),pp. 4626–4639.

[12] T.David,M.Ze-Qiang,M.Daniel,H.Amy-Joan,C.Matthew, *DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring*, Journal of Proteome Research,9(2008),pp. 3838–3846.

[13] F.Ari, *A ranking-based scoring function for peptide- spectrum matches*, Journal of Proteome Research,8(2009),pp. 2241–2252.

[14] N.Seungjin,B.Nuno,P.Eunok, *Fast multi-blind modification search through tandem mass spectrometry*, Molecular & Cellular Proteomics,11(2012),p. 012087.

[15] C.Hao,L.Chao,Y.Hao,Z.Wen-Feng,W.Long,Z.Wen-Jing,W.Rui-Min,N.Xiu-Nan,D.Yue-He,Z.Yao, *Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine*, Nature biotechnology,11(2018),pp. 1059–1061.

[16] N.Seungjin,K.Jihyung,P.Eunok, *MODplus: robust and unrestrictive identification of post-translational modifications using mass spectrometry*, Analytical Chemistry,91(2019),pp. 11324–11333.

[17] P.Ana-LS,O.Jose-TA,S.Gustavo,V.Ilka, *Label-free proteomic reveals that cowpea severe mosaic virus transiently suppresses the host leaf protein accumulation during the compatible interaction with cowpea*, Journal of Proteome Research,15(2016),pp. 4208-4220.

[18] N-Nathalie,T.Lucie,C.Cerina,A.Zuzanna,L.Joanna,G.Francois,B.Anne,E. Aleksander,A.Corinne,G.Ida-Chiara, *Impact of cystinosin glycosylation on protein stability by differential dynamic stable isotope labeling by amino acids in cell culture (SILAC)*, Molecular & Cellular Proteomics,16(2017),pp. 457-468.

[19] C.Liam,P.Daniela,L.Dennis,S.Ruth,T.Andreas, *Combination of bottom-up 2D-LC-MS and semi-top-down GelFree-LC-MS enhances coverage of proteome and low molecular weight short open reading frame encoded peptides of the archaeon Methanosarcina mazei*, Journal of Proteome Research,15(2017),pp. 3773–3783.

[20] P.Jillian,K.Anna,G.Harald,C.Ulisse,V.Matthijs,K.Manuel,B.Silvia,M.Marc, H.Craig,S.Brandon, *Chemosynthetic symbionts of marine invertebrate animals are capable of nitrogen fixation*, Nature Microbiology,2(2016),pp. 1-11.

[21] M.Clara,F.Bertrand,H.Maarten-LATM,P.Harriet,D.MichaelJ,L.KathrynS,N.BartM, *In-depth characterization of the tomato fruit pericarp proteome*, Proteomics,17(2017),p. 1600406.

[22] S.Gunnar,Me.David,S.Nesli-Ece,G.Anika,K.Sylvia,A.Georg, *Quantitative global proteomics of yeast PBP1 deletion mutants and their stress responses identifies glucose metabolism, mitochondrial, and stress granule changes*, Journal of Proteome Research,16(2017),pp. 504-515.

[23] H-Han,B.Kaspar,W.Jakob,Z.Fred,H.Yue,F.Mao,H.Bin,F.Yu,W.Abebe-Jenberie,L.Jianke, *Proteome analysis of the hemolymph, mushroom body, and antenna provides novel insight into honeybee resistance against varroa infestation*, Journal of Proteome Research,15(2016),pp. 2841-2854.

[24] C.Wojciech,L.Martina,P.Anne,N.Tuula,M.Sampsa, *Proteomic and bioinformatic characterization of extracellular vesicles released from human macrophages upon influenza A virus infection*, Journal of proteome research,16(2017),pp. 217-227.

[25] R.Daniel,A.Josef,M.Ulrike,R.Hermann,I.Till,a.Praveen-Kumar,T.Andrea,G.Cyprien,N.Pierre,S.Leif, *Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism*, Genome research,27(2017),pp. 298-299.

[26] C.Hao,L.Chao,Y.Hao,Z.Wen-Feng,W.Long,Z.Wen-Jing,W.Rui-Min,N.Xiu-Nan,D.Yue-He,Z.Yao, *Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine*, Nature biotechnology,36(2018),pp. 1059-1061.

[27] K.ThomasN,W.Max, *Semi-supervised classification with graph convolutional networks*, ArXiv,(2016).

[28] F.Zhengcong,W.Kaifei,C.Hao, *GameTag: A New Sequence Tag Generation Algorithm Based on Cooperative Game Theory*, Proteomics,20(2022),pp. 21-22.

[29] F.Zhengcong, *Novel Peptide Sequencing With Deep Reinforcement Learning*, 2020 IEEE International Conference on Multimedia and Expo (ICME),79(2022),pp. 1-6.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Apr 10, 2024
Accepted: May 27, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 160–168, DOI 10.12694/scpe.v26i1.3614

FAULT DIAGNOSIS OF CNC MACHINE TOOLS BASED ON SUPPORT VECTOR MACHINE OPTIMIZED BY GENETIC ALGORITHM

YONG WANG [*]AND CHUNSHENG WANG [†]

Abstract. To enhance the accuracy of CNC machine tool fault diagnosis, this study proposes an intelligent optimization method based on the combination of Particle Swarm Optimization (PSO) and Bacterial Foraging Algorithm (BFA), referred to as PSO-BFA. By simulating the local foraging behavior of bacteria, the PSO-BFA algorithm demonstrates characteristics of local convergence, replicability, and migratory properties during parameter selection, effectively improving the local optimization capability and fitness value of the model. This leads to faster convergence to the optimal solution in the fault data training process. The study utilizes a Deep Confidence Network (DCN) model, known for its strong adjustability of model structure, for training the fault feature set. The PSO algorithm is employed to search for the optimal value in the global range. Simulation data indicate that the PSO-BFA intelligent optimization method significantly outperforms traditional swarm intelligence methods in multi-fault diagnosis and classification, achieving the peak fitting value in fewer iterations.

Key words: CNC, PSO-BFA; local optimization; migratory properties

1. Introduction. Modern manufacturing organizations rely heavily on CNC (Computer Numerical Control) machine tools as their primary equipment due to their advanced capabilities in mechanical manufacturing technology, automatic control technology, signal control technology, and computer science. With increasing demands for precision and processing efficiency in machining products, the automation level of CNC machine tools has risen significantly [1, 2]. However, the structural complexity of these tools also increases the risk of failure, reducing their reliability and potentially leading to significant financial losses and safety hazards for operators. A sudden breakdown during high-speed operations can result in severe consequences, making real-time fault diagnosis and monitoring essential.

Fault diagnosis techniques for CNC machine tools have traditionally relied on detecting and analyzing fault signal characteristics, fuzzy reasoning, and other methods. However, as fault data sets grow larger, noise interference can lead to a decline in diagnostic accuracy, making it difficult to meet online monitoring requirements. To address these challenges, this paper proposes a hybrid intelligent optimization method combining Particle Swarm Optimization (PSO) and Bacterial Foraging Algorithm (BFA), referred to as PSO-BFA [3, 4].

The primary objective of this study is to enhance the accuracy and efficiency of fault diagnosis for CNC machine tools by leveraging the strengths of PSO and BFA algorithms [5, 6]. PSO is known for its global optimization capabilities, but it often struggles with local optimal solutions. BFA, on the other hand, mimics bacterial foraging behavior to achieve local optimization, enhancing the ability to adjust model parameters dynamically. By integrating these two algorithms, the PSO-BFA method aims to improve the local and global optimization capabilities of the fault data training model, leading to higher diagnostic accuracy and faster convergence to optimal solutions.

In this study, a Deep Confidence Network (DCN) model is selected for training the fault feature set due to its strong adjustability and effectiveness in handling large-scale data. The PSO algorithm is used to search for optimal values across the global range, while the BFA algorithm focuses on local optimization to avoid falling into local optima [7, 8]. The hybrid PSO-BFA approach allows for faster and more accurate parameter selection, reducing optimization time and improving the overall fault diagnosis process.

This paper is organized as follows: Section 2 describes the optimization of PSO-BFA swarm intelligence algorithm parameters. Section 3 presents the application of the PSO-BFA algorithm in online problem diagnosis

[*] Precision Manufacturing College, Suzhou Vocational Institute of Industrial Technology, China (happylifang806@163.com).
[†] Precision Manufacturing College, Suzhou Vocational Institute of Industrial Technology, China.

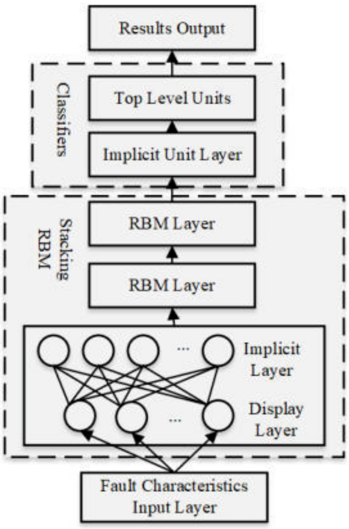


Fig. 2.1: CNC machine tool defect feature set training process.

for CNC machine equipment. Section 4 discusses the types and diagnoses of faults in CNC machine tools, along with an analysis of training samples for fault phenomena and diagnostic convergence effects. Finally, the study's conclusions and future research directions are outlined.

By improving the fault diagnosis techniques for CNC machine tools through the PSO-BFA intelligent optimization method, this research aims to significantly enhance the reliability and efficiency of CNC machine tool operations, ultimately contributing to the advancement of modern manufacturing technologies [9].

2. Optimizing PSO-BFA swarm intelligence algorithm's parameters. In this paper, we select the deep confidence network model with stronger adjustability of model structure among existing large-scale fault data training methods, such as production adversarial network model [10, 11, 12, 13], convolutional neural network model [14], and deep confidence network model [15]. The main component of fault feature set training is model parameter optimization and adjustment, which enhances the fault set training model's capacity for both local and global optimization and, eventually, raises the diagnostic accuracy of machine tool failures. By combining the benefits of the BFA and PSO group intelligence algorithms, the PSO-BFA intelligent optimization algorithm selects the model parameters more quickly, accurately, and with a shorter optimization time. There is a reduction in the amount of time needed for optimization, more accuracy, and speedier parameter selection. The structural design of the deep confidence network model is depicted in Fig. 2.1. The deep confidence network is composed of multiple RBMs stacked (restricted Boltzmann machines) that process the bottom layer input data [16]-[17], intermediate hidden layer data training, and top layer unit output training results.

Due to its high global optimization capabilities, the PSO algorithm has an advantage in the parameter optimization and training model selection processes. Assuming that each solution in the D-dimensional space of the defect feature set is a particle, the total number of n , as well as the starting velocity and position of the i-th particle, are represented as follows:

$$\left\{\begin{array}{l} l_i=\left(l_{i1}, l_{i2}, \cdots, l_{iD}\right), i=1,2, \cdots, n; \\ v_i=\left(v_{i1}, v_{i2}, \cdots, v_{iD}\right), i=1,2, \cdots, n_{\circ} \end{array}\right. \tag{2.1}$$

The th particle's inertia weight is ω , and the learning factors are κ_1 and κ_2, both in the interval [0, 1]. These factors initialize the population particles' location and velocity, of which the global ideal position is L_g , and the individual optimal position is L_i.The particles move in space at a specific speed, and in order to achieve the global optimization in the range, they dynamically adjust their own velocity and the present occupied position based on their own and other people's movement experiences. The following describes the procedure used to update the ith particle's position and velocity at the $k+1$st moment:

$$\left\{\; l_i^{k+1} = l_i^k + v_i^{k+1}\right. \tag{2.2}$$

where the model's random number is represented by rand. According to the particle swarm population size to determine the proper fitness function, calculate the fitness value of the particles in the swarm and compare it with the global optimal extreme point.The BFA algorithm is combined with the traditional PSO algorithm in this paper to optimize and enhance the parameters of the confidence network model. While the PSO algorithm can achieve optimization in the global range, it is prone to settling into the local optimal solution.The foundation of the BFA algorithm is the idea of mimicking the foraging behavior of bacteria in order to accomplish local optimization. The basic idea behind the BFA algorithm is to mimic bacterial foraging in order to accomplish the goal of local optimization. The PSO algorithm's population particles are attributed to the foraging role of bacteria, meaning that they possess the traits of migration, replicability, and convergence.Individual bacteria are capable of local optimization seeking, which allows them to swim to the local enriched zone and modify the fitness value target in real time.In the PSO-BFA algorithm, the jth convergence operation process of particle $i$is represented by$\eta_i\left(j,k,l_j\right)$, the number of replication operations is indicated by$k$, and particle i's current fitness value is $J_i\left(j,k,l_j\right)$.Assuming that particle$i$'s movement direction and moving step size are represented by the symbols$\varphi(i)$and$c(i)$, we can represent particle i's single convergent operation as follows:

$$\left\{\; \eta_i\left(j+1,k,l_j\right) = \eta_i\left(j,k,l_j\right) + \varphi(i) \times c(i)\right. \tag{2.3}$$

The PSO-BFA algorithm's population particles have bacterial pheromone detecting capabilities, which allows them to detect the nutritional data that other particles near individual i are carrying. One benefit of the PSO-BFA algorithm is its localized sensing capacity. The benefit of local sensing is that it can maintain a healthy spacing between particle people and prevent population members from being very dense in one area. By determining the distance between all nearby particles—that is, all particles in the population—particle i gathers and sends nutritional information to the outside world. This mechanism is explained as follows:

$$\zeta\left(\tau,\eta_i\left(j,k,l_j\right)\right) = \sum_{i=1}^{n}\left[-\zeta_{att}\exp\left(-w_{att}\sum_{k=1}^{m}\tau_m - \tau_m^i\right)^2\right] + \sum_{i=1}^{n}\left[-\zeta_{rep}\exp\left(-w_{rep}\sum_{k=1}^{m}\tau_m - \tau_m^i\right)^2\right] \tag{2.4}$$

where τ is the composite pheromone of particle transfer information, which comprises the current particle's position, direction, and velocity, among other details; The gravitational depth between particles is represented by ζ_{att}, the repulsive depth by ζ_{rep}, the gravitational width by w_{att}, and the repulsive width by w_{rep}. The PSO-BFA technique is able to realize the local optimization of the adjustment of the training parameters of the defective dataset of the machine tool of the deep confidence network model by utilizing the group sensing mechanism between particles [18]-[19].

3. Online problem diagnosis for CNC machine equipment. The PSO-BFA algorithm is used to locally search all population particles for a certain amount of time. The natural evolutionary law of organisms states that certain individuals are removed if they cannot locate adequate food or are in an unfavorable location [20]. A population's convergent behavior—in which the total individual fitness of the population is used to determine its activity level—must periodically be confirmed, even if healthy individuals will replicate to maintain the number of active individuals in the population. The BFA technique allows for local optimization and prevents parameter selection from falling into local optimal solutions, whereas the standard PSO algorithm excels in global optimization. The formula for the movement speed of the population particles defines the traveling direction of the entire population as well as the optimization process. The moving speed V of the population at the $k+1$-th moment is given as follows:

$$V = \omega v_i^{k+1}\zeta\left(\tau,\eta_i\left(j,k,l_j\right)\right) + \kappa_1 rand()\left(L_i^k - l_i^k\right) + \kappa_2 rand()\left(L_g^k - l_i^k\right) \tag{3.1}$$

Following the implementation of the BFA algorithm for local step update, the most recent position is indicated as follows:

$$L_g^{k+1} = V + L_{g^\circ}^k \tag{3.2}$$

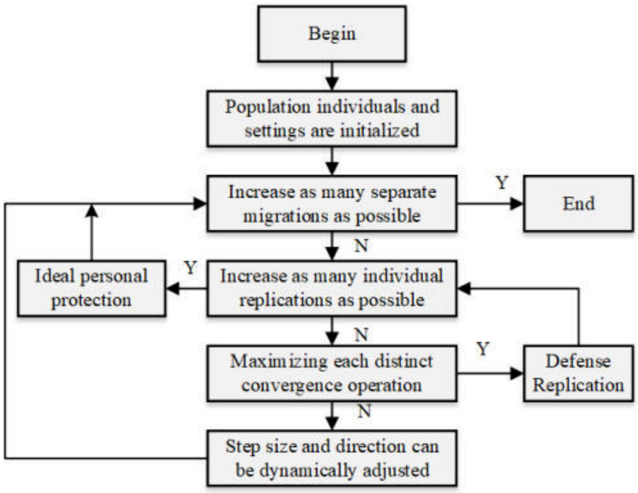

Begin
Population individuals and settings are initialized
Increase as many separate migrations as possible
Y
End
N
Ideal personal protection
Y
Increase as many individual replications as possible
N
Maximizing each distinct convergence operation
Y
Defense Replication
N
Step size and direction can be dynamically adjusted

Fig. 3.1: Local parameter optimization of the PSO-BFA algorithm procedure.

Throughout the population migration process, the PSO-BFA algorithm will be assigned a probability value p_0. In order to prevent falling into the optimal solution during the local optimization, the migration operation of the population individuals will only be carried out at this point if the probability of the random number selection is lower than p_0 during the individual migration.In order to prevent the population from growing as a result of individual elimination and situations involving individual reduction and return, the PSO-BFA algorithm also considers the particle swarm global search optimization problem during local optimization. This ensures both the accuracy of global optimization and the velocity of movement. Fig.3.1 illustrates the specific parameter optimization process[21].

The individual convergence operation's goal is to filter out the best individuals from the current population. The immune replication link is the essential component of the fault data training model's parameter optimization process [22]. In order to determine whether or not the global optimum has been reached, the following steps are taken: first, the fitness value of each individual is calculated; second, the excellent individuals are replicated in order to maintain the population's size; and third, the immunoreplicated population's overall fitness is optimally calculated. The traditional BFA methodology is optimized by the use of immunoreplication. In the standard BFA mode, the individuals are sorted according to their fitness values, and the locations of the reproduced individuals are the same as the original ones, i.e., the placements of the people are not optimized, but only the fitness values of the individuals are improved. In contrast, in the immunological replication mode, the replication of individuals is based on convergence, and the particle population individuals have strong foraging capacity following replication and high-frequency mutation[23]. To fulfill the goal of individual local optimization searching, its place within the population is further optimized. The replication target was selected from among the elite individuals with the highest degree of adaptability[24]; n/2 replicated individuals in total. Here, the individual adaptability values were computed for every member of the population following replication and mutation. Start the cycle processing for individual migration number maximization, replication number maximization, and individual convergence operation maximization. Once the BFA algorithm has been improved to an optimal level, each particle swarm individual is given the ability to search for optimization, which includes both direction and step length control. This means that each particle swarm member will swim towards the local nutrient-rich region until its fitness value stops increasing, at which point its local optimization-seeking will have found the best solution.

$$\left\{ \eta_{\max}\left(j+1,k,l_n\right)=\eta_n\left(j,k,l_n\right)+\varphi(i)\times c(i) \right. \tag{3.3}$$

At this point, the population particles' local convergence operation reaches its peak $\eta_{\max}\left(j+1,k,l_n\right), \varphi(i)\times c(i)$ value and begins to drift toward zero. When the individual convergence operation reaches the optimum,

Table 4.1: A portion of the set related to the failure of CNC machine tools.

No.	Warning Indications of Failure	No.	Warning Indications of Failure
1	Non-functioning spindle motor	6	There is no spindle lubricant circulation
2	Erratic motor speed	7	The servo motor is broken
3	Rotation of the spindle stops	8	axis tremor when cutting
4	Heating of the spindle	9	harm to electrical parts
5	Upon severe cutting force, the spindle stops	10	Electrical parts can burn out, smoke and catch fire, or overheat.

Table 4.2: A portion of the set related to the failure of CNC machine tools.

No.	Warning Indications of Failure	No.	Warning Indications of Failure
1	The enable signal is not connected correctly	8	Ball Screw Sub-Gap
2	The knife frame itself failing	9	Overly
3	faulty connection in the electrical wiring	10	Interference
4	Inadequate motor performance or personal error	11	Inadequate lubrication
5	Zero switch lacks sensitivity	12	Connecting wires and the encoder
6	unreasonable parameter configuration	13	Negative
7	Insecure connections	14	Overload

the local fitness value of the individual is no longer increased on the elite individuals to replicate and mutate to better optimize the population. If the individual convergence operation is not maximized, to continue adjusting the step size and direction of the individual until the local optimum is reached.The optimized population's convergence performance is finally confirmed. Population A_0 has n active population members in total. The immune-immune space geometry for any given initial state of the population is I^*. There are N optimal solutions in total for the population, and the set of optimal solutions is B_N^*. This means that the numerical training parameter search for fault characteristics finds the optimal solution both locally and globally when the following probability distribution conditions are met:

$$\lim P\left[\eta(N)1 \mid \eta(0)=A_0\right]=1 \tag{3.4}$$

The population members in the local and global scope are in a state of quasi-optimality when the algorithm cyclically replicates and adjusts for convergence. At this point, the PSO-BFA algorithm's optimal parameter selection also tends to converge, allowing the depth of the confidence network to be utilized for fault feature recognition and data training.

4. Types and diagnoses of faults in CNC machine tools. The machinery industry has made extensive use of CNC machine tools, and as a result, maintenance and repair work on these tools has grown in importance. In order to improve the maintenance efficiency of CNC machine tools, maintenance level and fault diagnosis rate, row of CNC machine tool system failure is one of the important tasks in the occurrence of CNC machine tool failure, the failure of the triggering factors for judgment and assessment, so as to facilitate the accurate determination of the location of the fault occurred in a timely and effective way to take appropriate fault diagnosis measures for fault repair [22]. Table4.1 and 4.2 illustrate frequent failure phenomena and triggers for CNC machine tools.

4.1. Analysis of training samples for fault phenomena. The experimental apparatus is a CNC lathe, model number CAK6150. Common fault indications and root causes are categorized, and the likelihood of faults occurring is mined and fed into an SSA-BP neural network. Owing to the wide range in fault frequency, the diagnostic method generates a grade for the fault indications that is classified as high, medium, or low. The particular outcomes are displayed in Fig. 4.1. The cause-and-effect relationship between common faults and machine tool processing repetition should be understood in light of the information presented in Fig. 4.1. As an illustration, consider the spindle frequent defects.

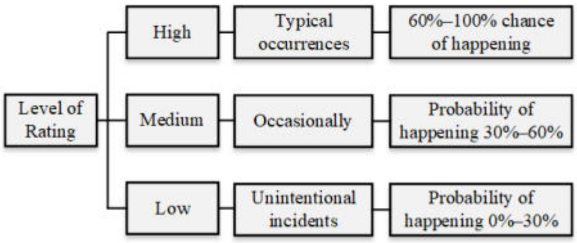

Level of Rating
High
Typical occurrences
60%–100% chance of happening
Medium
Occasionally
Probability of happening 30%–60%
Low
Unintentional incidents
Probability of happening 0%–30%

Fig. 4.1: Chance of a fault occurring.

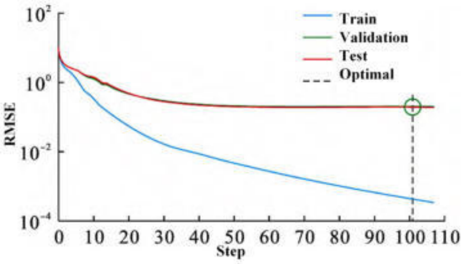


Fig. 4.2: BP Neural Network Convergence.

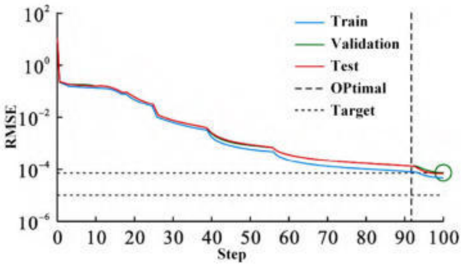


Fig. 4.3: SSA-BP neural network convergence.

4.2. Diagnostic Convergence Effect Comparison. The BP neural network's convergence is depicted in Fig. 4.2. The substantial simulation error, poor iterative convergence impact, and large error oscillation phenomenon of the conventional BP network on frequent failures of CNC machine tools.

Fig. 4.3 illustrates the SSA-BP neural network convergence after it was trained using MATLAB software and retested utilizing the CNC machine tool defect data that occurred under real-world operating conditions as the network sample set. How the SSA-BP neural network can achieve a 100 iteration effect. Its convergence speed is faster and its convergence curve is smoother than that of the regular BP neural network.

Fig. 4.4 shows the prediction fault output error of the SSA-BP neural network. The SSA-BP neural network prediction error, which is 2.29%. This is quite near to the theoretical output, which is the expected effect corresponding to the theoretical deviation. In comparison to a typical BP neural network, the prediction fault output error of the SSA-BP neural network is smaller.

4.3. Defect Recognition. The moment of occurrence of each transient in the vibration signal is converted into a series of sparse representation coefficients using this approach of sparsely representing the vibration signal. This allows for the identification of transients and the diagnosis of automatic tool change system defects. Fig. 4.5 displays the appropriate ideal dictionary atom.

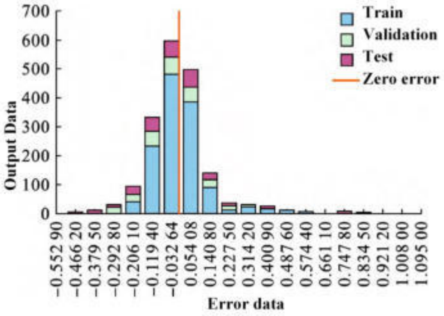


Fig. 4.4: A neural network with SSA-BP predicts fault output error.

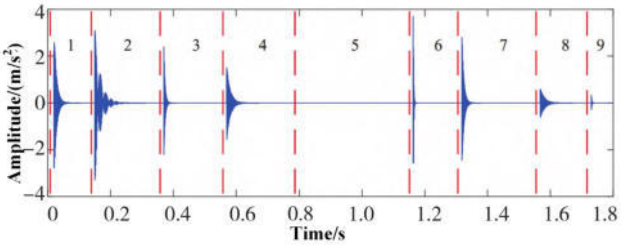


Fig. 4.5: Ideal dictionary atoms.

Fig. 4.6 displays the results of the sparse representation achieved by the proposed method in this chapter. The tool-changing timing diagram's results are largely consistent with the occurrence moments of each transient, which can be obtained from Fig. 4.6. The accuracy of the extracted time intervals of the neighboring transients is very high, as evidenced by the mere 0.7% relative error. Fig.4.6demonstrates how the technique can successfully pinpoint each tool change vibration signal transient's moment of occurrence and extract the properties of nearby transients.

5. Conclusion. This research suggests a PSO-BFA-based intelligent optimization technique. In order to enhance the adaption value and the local optimization capability of model parameters, the BFA algorithm is presented to replicate the local foraging behavior of bacteria. The fault data training model chooses a deep confidence network model with customizable scale. The suggested approach performs noticeably better than the conventional technique in terms of multi-fault classification and diagnosis capacity, according to simulation findings.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

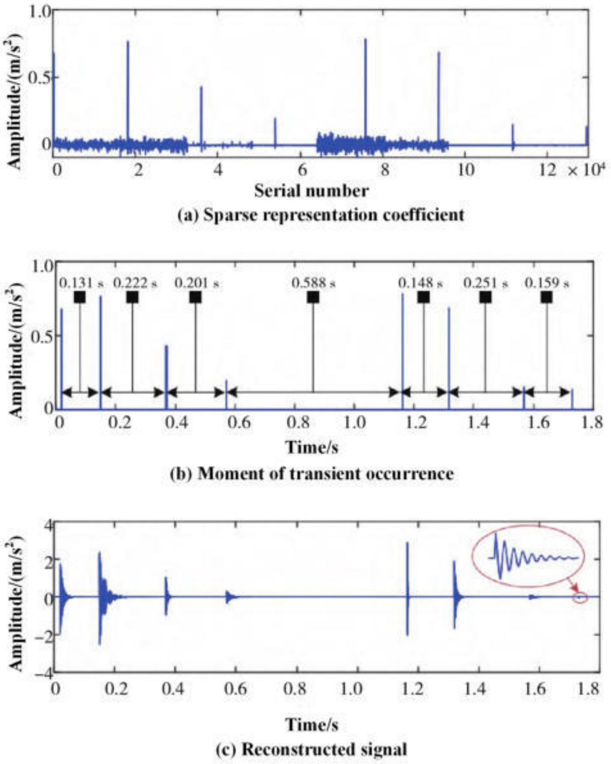
REFERENCES

[1] Jaafar, R., Ab Wahab, N., Bidin, B. B., Johari, M. N. M. B., Kaliswaran, A., Mogal, L., Fedileh, M. I.,*An Interim Study of Integrating Spindle, Laser, and Plotter in a CNC Router Machine*. Engineering Science Letter, 1(02),(2022), 55-60.

[2] Huang, S., Lu, N., Jiang, B., Simani, S., Li, R., Huang, B., Cao, J.,*Fault propagation analysis of computer numerically controlled machine tools*. Journal of Manufacturing Systems, 70,(2023), 149-159.

[3] Wang, J., Yin, W., Gao, J.,*Cases Integration System for Fault Diagnosis of CNC Machine Tools Based on Knowledge Graph*. Academic Journal of Science and Technology, 5(1),(2023) 273-281.

[4] Fahrizal, F., Aslam, M. F., Anwar, N., Isminarti, I., Fitriati, A.,*Design of Styrofoam Cutting Machine Based on CNC 2 Axis Using Hot Wire*. Journal of Computer Engineering, Electronics and Information Technology, 1(2), 51-58.



(a) Sparse representation coefficient

(b) Moment of transient occurrence

(c) Reconstructed signal

Fig. 4.6: Outcomes of the vibration signal's sparse representation for a typical tool change cycle.

[5] hairol mizzam Haris, M., Zakaria, R. B., Dan, N. B., *Design of Mould Vacuum Thermoforming Machine using CNC Machine.* Journal on Technical and Vocational Education, 8(2),(2023) 36-43.

[6] Gu, D., Xu, Z., Zhong, Y., Li, Q., Long, Z.,*Reliability allocation method of comprehensive weight computer numerical control machine tool based on failure correlation and factor correlation.* Quality and Reliability Engineering International, 39(8),(2023) 3285-3302.

[7] Soori, M., Arezoo, B., Dastres, R.,*Machine learning and artificial intelligence in CNC machine tools, A review.* Sustainable Manufacturing and Service Economics,(2023),100009.

[8] Liu, W., Zhang, S., Lin, J., Xia, Y., Wang, J., Sun, Y. *Advancements in accuracy decline mechanisms and accuracy retention approaches of CNC machine tools: a review.* The International Journal of Advanced Manufacturing Technology, 121(11-12), (2022),7087-7115.

[9] Bin, Z. H. U., Liping, W. A. N. G., Jun, W. U., Hansong, L. A. I.,*Reliability modeling and evaluation of CNC machine tools for a general state of repair.* Journal of Tsinghua University (Science and Technology), 62(5), (2022),965-970.

[10] Jung, S., Kim, M., Kim, B., Kim, J., Kim, E., Kim, J., ... Kim, S. *Fault Detection for CNC Machine Tools Using Auto-Associative Kernel Regression Based on Empirical Mode Decomposition.* Processes, 10(12),(2022), 2529.

[11] Amaya-Toral, R. M., Piña-Monarrez, M. R., Reyes-Martínez, R. M., de la Riva-Rodríguez, J., Poblano-Ojinaga, E. R., Sánchez-Leal, J., Arredondo-Soto, K. C.,*Human–machine systems reliability: A series–parallel approach for evaluation and improvement in the field of machine tools.* Applied Sciences, 12(3),(2022),1681.

[12] Shicong, P., Guocheng, W., Fuqiang, T.,*Design and realization of CNC machine tool management system using Internet of things.* Soft Computing, 26(20), (2022),10729-10739.

[13] Zhang, Z., Yang, Y., Li, G., Qi, Y., Yue, C., Hu, Y., Li, Y.,*Machining accuracy reliability evaluation of CNC machine tools based on the milling stability optimization.* The International Journal of Advanced Manufacturing Technology, 124(11-12),(2023), 4057-4074.

[14] Mechta, A., Slamani, M., Zaoui, M., Mayer, R., Chatelain, J. F.,*Correlation assessment and modeling of intra-axis errors of prismatic axes for CNC machine tools.* The International Journal of Advanced Manufacturing Technology, 120(7-8),(2022) 5093-5115.

[15] Zhu, M., Yang, Y., Feng, X., Du, Z.,Yang, J., *Robust modeling method for thermal error of CNC machine tools based on random forest algorithm.* Journal of Intelligent Manufacturing, 34(4),(2023) 2013-2026.

[16] Iqbal, M., Madan, A. K.,*CNC machine-bearing fault detection based on convolutional neural network using vibration and acoustic signal.* Journal of Vibration Engineering & Technologies, 10(5), (2022),1613-1621.

[17] Nebelung, N., de Oliveira Santos, M. D., Helena, S. T., de Moura Leite, A. F., Canciglieri, M. B., Szejka, A. L.,*Towards Real-Time Machining Tool Failure Forecast Approach for Smart Manufacturing Systems.* IFAC-PapersOnLine, 55(2),(2022), 548-553.

[18] Kuo, C., Lin, X., Yeh, T.,*Working towards the minimum surface damages and failure analysis of Joule heat effects in manufacturing diamond cutting tools.* Engineering Failure Analysis, 152,(2023), 107432.

[19] Feng, C., Yang, Z., Chen, C., Guo, J., Tian, H., Meng, F.,*Quantitative evaluation method for machining accuracy retention of CNC machine tools considering degenerate trajectory fluctuation.* Journal of Mechanical Science and Technology, 36(6), 3119-3129.

[20] Wu, Y., Yang, Z., Wang, J., Chen, X., Hu, W.,*Optimizing opportunistic preventive maintenance strategy for multi-unit system of CNC lathe.* Journal of Mechanical Science and Technology, 36(1),(2022),145-155.

[21] Han, C., Lin, T.,*Reliability evaluation of electro spindle based on no-failure data.* Highlights in Science, Engineering and Technology, 16,(2022), 86-97.

[22] Dai, Y., Tao, X., Li, Z., Zhan, S., Li, Y., Gao, Y. ,*A review of key technologies for high-speed motorized spindles of CNC machine tools.* Machines, 10(2),(2022),145.

[23] Jingchun Zhou, Qian Liu, Qiuping Jiang, Wenqi Ren, Kin-Man Lam, Weishi Zhang.,*Underwater image restoration via adaptive dark pixel prior and color correction.* International Journal of Computer Vision, 2023. DOI :10.1007/s11263-023-01853-3.

[24] Ali, J., Shan, G., Gul, N., Roh, B. H.,*An Intelligent Blockchain-based Secure Link Failure Recovery Framework for Software-defined Internet-of-Things.* Journal of Grid Computing, 21(4),(2023), 57.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Apr 1, 2024
Accepted: May 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 169–179, DOI 10.12694/scpe.v26i1.3628

INVESTIGATION INTO THE OPTIMISATION OF COLD CHAIN LOGISTICS DISTRIBUTION PATHS USING THE HYBRID ANT COLONY METHOD

WEIWEI XU*

Abstract. China's cold chain logistics market has been growing quickly in recent years. Cold chain logistics helps minimize food loss and waste during transit in addition to meeting people's need for fresh food. As the idea of "green logistics" has gained traction, we created a better ant colony algorithm with a multi-objective heuristic function to address the issue. Specifically, we combined the A^* algorithm with the ACO algorithm to address the issue of insufficient pheromone in the early stages of the ACO algorithm, and the resulting improved multi-objective ACO algorithm was able to solve the vehicle path distribution problem with a multi-objective optimisation model more successfully than the traditional ACO algorithm, yielding more Pareto efficient solutions. Ultimately, simulation studies demonstrate that the distribution paths produced by the multi-objective model and algorithm presented in this paper can concurrently optimize for lowering distribution costs, cutting carbon emissions, and raising customer satisfaction, ultimately resulting in a more ecologically friendly and greener distribution solution.

Key words: vehicle path problem; cold chain logistics; energy saving and emission reduction; hybrid ant colony algorithm

1. Introduction. With the continuous improvement of the quality of life, people's demand for fresh green fresh products is also increasing [1]. In order to meet the market demand and promote the development of enterprises, cold chain logistics enterprises have invested a large number of vehicles in the transport link, but because the fuel consumption and carbon emissions generated in the process of cold chain logistics and distribution are far more than that of ordinary logistics, the pressure on enterprises to reduce the operating costs and the negative impact of automobile exhaust on the environment is also increasing. According to the survey of China Logistics and Purchasing, nearly half of the logistics enterprises' fuel expenses account for more than 40% of the transport costs [2]. According to the statistics of the World Resources Institute, the carbon emissions of the transport industry account for 20% of the total global emissions [3]. Facing the double pressure of economy and environment, energy saving and emission reduction is especially important for cold chain transport enterprises [4]. By promoting energy saving and emission reduction and controlling the amount of fuel consumption, enterprises not only compress the cost of fuel consumption on the one hand, but also, with the government of China continuously accelerating the full implementation of the carbon emissions trading system [5], it is likely to reduce the operating cost of carbon trading for the enterprise in the near future and increase the profit of the enterprise, which is conducive to the development of the enterprise. On the other hand, because carbon emissions depend on the amount of fuel consumption, while reducing the use of fuel resources also reduces the carbon pollution caused by the environment, in line with the concept of green logistics development [6]. Therefore, it is very important to add energy saving and emission reduction into the cold chain logistics vehicle path problem in order to seek a win-win situation between economy and environment for the rapid development of cold chain logistics enterprises.

Environment-friendly society is a social system that consists of environment-friendly technologies, products, enterprises, industries, schools, communities, etc., aiming at man and nature, based on environmental carrying capacity, following the law of nature as the core, and at the same time advocating environmental culture and ecological civilisation, and pursuing the coordinated development of economy, society and environment [7]. Environment-friendly society emphasizes the pursuit of harmony between man and nature as the goal, based on the carrying capacity of the environment, to follow the laws of nature as the core, to achieve the environment and the economic and social comprehensive and coordinated sustainable development of social relations, and

*School of Economics and Management, Zhejiang Ocean University, Zhoushan 316022, Zhejiang, China (chenyunzhou1990@zjgsdx.edu.cn).

to achieve the transformation of the industrial civilization from predatory, conquering and polluting to a coordinated, restorative and constructive ecological civilization[8]. Economic benefits frequently take precedence over energy consumption and environmental effects in conventional logistics and distribution channel design [9]. Green logistics calls on us to incorporate energy conservation, emission reduction, and other environmental protection components into the path optimization issue in order to minimize the adverse effects on the environment while maintaining economic advantages [10]. As a result, researching path optimization in cold chain logistics has a lot of application and aids businesses in developing green logistics in a sustainable manner.

There are two main issues: 1) In order to reduce the overall distribution cost, some scholars have focused the cold chain logistics route optimisation problem on vehicle path optimisation under static networks [11]. 2) The ACO algorithm is widely used and has strong parallelism and robustness [12].

Scholars have categorized the VehicleRouting issue (VRP) as an NP-hard issue, meaning that solving this kind of problem with precise mathematical analytical techniques is challenging [13]. A number of academics have successfully solved VRP issues using heuristic methods [14]. This work uses the ant colony method to solve the model since it is resilient, parallel, and commonly utilized to tackle VRP issues [15]. The absence of a route pheromone in the early stages of the ant colony algorithm can easily result in blind search, which raises the number of convergence excessively high [16]. In order to tackle this issue, the model in this paper is proposed to be solved using a hybrid ACO algorithm in conjunction with the A^* algorithm. The route information is initialized to minimize the number of convergence times and decrease the convergence time of the ACO algorithm based on the optimal solution found by the A^* algorithm. In addition, the heuristic factor and transfer probability are enhanced in accordance with the research findings in this work, making the hybrid ant colony algorithm better suited to the issues this article will be studying.

2. Problem description. Researchers both domestically and internationally have been delving deeply into the topic of cold chain logistics in recent years. The influence of the number of delivery vehicles, client demand, and delivery time on the overall cost of cold chain distribution was investigated in study [17], which also used a genetic algorithm to solve the cold chain distribution route optimization model with tight time window limitation. Improved results were obtained by using a heuristic approach to solve a dynamic multi-objective vehicle route optimization model, as investigated in study [18].

The logistics route issue with time windows was investigated in study [19]. The ant colony algorithm was enhanced by modifying the pheromone for model solution, and the method's efficacy was confirmed by case analysis and comparison. Research [21] employed the enhanced simulated annealing approach to analyze the variation in distribution costs while taking into account the scenario of heterogeneous fleet based on the conventional route model.

When building a cold chain logistics path model, research [22] takes into account the variety of client demand types. An ant colony algorithm and clustering techniques are used to solve the model. Study [23] established a mathematical model under the constraints of vehicle load and customer demand, took into account the cost of cargo damage in the distribution process in relation to the perishability of fresh products, designed a traditional genetic algorithm in accordance with the model, and solved the model with an improved genetic algorithm. Some academics have focused on the study of low-carbon cold chain logistics as a result of the government's increasing demands for environmental preservation and sustainable development as well as people's growing awareness of energy conservation and emission reduction. In order to create a cold chain inventory model, study [25] considers carbon emission as a cost element. It then investigates the best inventory option given the limitations of cost and carbon emission, and it creates a precise algorithm to solve the model. Research [26] converted carbon emissions into economic costs to create a path optimisation model with fuzzy time; to improve results, the model was solved using an ant colony algorithm; research [27] created a path optimisation model that took into account both carbon emissions and customer time windows simultaneously, and it was solved using a heuristic algorithm.

In this work, we address the optimization problem of cold chain logistics paths for a single distribution center, where trucks originate from one center and return there upon fulfilling the delivery task of each customer's fresh goods. In actuality, cold chain logistics companies don't construct a lot of distribution centers because of budgetary constraints.

This paper's primary goal is to create a distribution strategy for a distribution center based on the resources at hand, accounting for variables like product demand, access time, and maximum transportation capacity. The aim of this initiative is to reduce the overall expenses incurred by the vehicle in relation to ”wasteful,” ”green,” ”indirect,” ”indirect,” ”road damage,” and ”soft” costs related to incarceration, all while fulfilling the necessary requirements for product distribution between customs points and the district center. From the standpoint of reducing emissions and conserving energy, this is done.

Below are the particular presumptions:

- The distribution center has a enough number of identically equipped refrigerated trucks to fulfill consumer demand for fresh produce delivery; also, the vehicles are capacity-limited, meaning that demand at each customer site won't surpass the vehicle's maximum capacity;
- Every customer point's location, the need for fresh goods, the amount of time needed for servicing, and the window for delivery are all known;
- A single reefer can transport goods to several locations, but it can only depart from and arrive at each location once; only one reefer truck is scheduled to deliver to the same customer location and can ensure that the service meets the needs of the customer location;
- A fine shall be paid by the business to the refrigerated truck operator if the vehicle carrying fresh product to a customer site is not delivered within the window of time that was arranged with the customer site;
- During the delivery service for a customer point, the refrigerated truck does not load or unload any items; instead, it just loads and unloads recently acquired items that come from the service for the client point; it does not accept other delivery services. The refrigerated truck returns to the distribution center when the distribution task is finished;
- The drivers employed in cold chain distribution are all subjected to the same rigorous training and technical experience, and subjective variables do not affect the fuel usage.

3. Examining the Model's Known Parameters and Variables. The traditional ant colony algorithm uses a positive feedback mechanism to continuously converge during the search process by calculating the transfer probability based on the distance between paths and the concentration of pheromones. However, this algorithm has the drawback of prematurely settling into the local optimum.This work limits the range of pheromone concentration, performs global pheromone updating with a cycle of 20 iterations in the algorithm design, and dynamically increases pheromone concentration in order to prevent the ant colony algorithm from entering a state of local optimality and enhance the accuracy of the solution. The time restriction is now taken into account by the ants in addition to distance when they move to the next node, thanks to a revision of the heuristic algorithm. The multi-objective optimization of the Pareto solution set is realized by setting the parameter of the pseudo-random proportional action selection rule such that the ants have a high chance of selecting the best path and ultimately approaching the optimal solution.

3.1. Recognized parameters. The cold chain logistics path optimization model that is the subject of this paper's study has the following known parameters.

N: Every client that the distribution center must service;

K: The total amount of cars at the district center that are needed for distribution;

f_k: The one-time fee for utilizing a refrigerated car k;

$fuel$: the quantity of fuel produced during the distribution of vehicles;

a: The distribution vehicles' refrigerant usage component during the transportation phase;

b: Distribution vehicles' refrigerant consumption coefficient throughout the loading and unloading process;

q_i: The level of fresh product demand at consumer point i;

p: The cost per unit of fresh product that the delivery vehicle transports;

∂_1: The freshness degradation coefficient of fresh goods while the delivery truck is in motion;

∂_2: The rate at which fresh goods loses freshness while being loaded and unloaded from the distribution vehicle;

t_i^k: The moment at which delivery vehicle k reaches client location i;

t_o^k: The moment at which delivery vehicle k leaves the distribution facility;

T_i: The duration of service for the distribution vehicle at customer point i;

Q_{ij}: The weight must move in order to get straight from customer point i to customer point j;

Q: The vehicle's maximum load capacity;

ε_1: A penalization element in case the delivery vehicle reaches the client's address prior to the mutually agreed-upon maximum time range;

ε_2: The penalty factor in case the delivery vehicle reaches the customer's location beyond the prearranged time window's lower bound.

3.2. Variable analysis.

3.2.1. Analysis of decision variables. In order to make the model analysis easier, the client point is represented by the letters i, j ,$i, j = 1, 2, 3, ..., (N)$ as well as the distribution facility receiving the number 0.

The following values are accepted for the decision variable x_{ijk}.

In the event that vehicle k travels directly along path i from customer point i to customer point j, j equals 1, otherwise x_{ijk} equals 0.

When car k fulfills i 's delivery order and delivers the products at customer point i, the response is 1, while in the other case, it is 0.

3.2.2. Cost Variable Analysis.

(1) Fixed expenses during the distribution of cars (C_1). A certain fixed cost, usually related to the number of activated refrigerated vehicles, is needed to activate the refrigerated vehicles that are used to provide distribution services to each customer point from the distribution center. These costs include vehicle depreciation, maintenance, and driver compensation. This is demonstrated by Eq.3.1:

$$C_1 = \sum_{k=1}^{K} \sum_{j=1}^{N} x_{0jk} f \tag{3.1}$$

where the value is 0 otherwise and 1 when the distribution center turns on the refrigerated truck K.

(2) The deployment of vehicles involves green costs (C_2).. The fuel costs incurred by the business for the fuel used in the transportation of refrigerated vehicles and the environmental costs incurred for the carbon emissions produced during transportation that lead to contamination of the environment are considered the "green costs" in this study.

The vehicle load and fuel consumption rate have a specific linear relationship. The typical driving unit distance fuel consumption rate of a reefer truck is ρ_0 when the truck is empty; once the vehicle is completely loaded, the normal driving unit distance fuel consumption rate is ρ_* . Eq.3.2 illustrates the typical driving unit distance fuel usage when the reefer truck is carrying goods weighing M .

$$\rho(M) = \rho_0 + \frac{\rho_* - \rho_0}{Q} M \tag{3.2}$$

where Q is the vehicle's maximum load capacity.

Fuel consumption from client location i to client location j is shown in Eq.3.3.

$$\rho\left(Q_{ij}\right) d_{ij} \tag{3.3}$$

Eq.3.4 can be used to calculate the fuel usage during the entire distribution process once the truck has finished serving every consumer point.

$$fuel = \sum_{k=0}^{K} \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ijk} \rho\left(Q_{ij}\right) d_{ij} \tag{3.4}$$

Eq.3.5 illustrates the cost of fuel used during the entire distribution process.

$$C_{21} = cfuel \tag{3.5}$$

where c is the oil price.

Fuel usage × carbon dioxide emission factor = carbon emissions, according to Ottmar's analysis. These two variables have a particular linear relationship. Eq.3.6 illustrates the environmental cost of the entire distribution process:

$$C_{22} = wfuel \tag{3.6}$$

for which the coefficient of carbon emissions is w .

Eq.3.7 illustrates the green cost associated with transporting chilled vehicles, as fuel costs and environmental costs follow a particular linear relationship with fuel usage.

$$C_2 = C_{21} + C_{22} = (c+ \downarrow \omega)fuel = \gamma fuel \tag{3.7}$$

where the coefficient of green costs is γ .

(3) The price of refrigeration for vehicles being distributed (C_3). The price of the refrigerant used to keep the temperature inside the car compartment is typically what is considered the refrigeration cost during the distribution process of refrigerated vehicles. This calculation does not account for the fuel used for refrigeration because the fuel used for refrigeration is already included in the green cost. The heat load that the car experiences while it is traveling, the degree of vehicle cracking, the heat transfer rate, the area of the compartment that receives solar radiation. The cost of refrigeration during vehicle transportation can be roughly estimated as positively correlated with the vehicle operating time because, according to assumption Eq.3.1, the distribution center's vehicles are all of the same type, have similar compartment parameters and levels of deterioration, and have relatively stable internal and external environments during driving. With the equipment already unloaded and disposed of, the distribution service can be fulfilled by simply opening the premises, i.e., the deployment costs and the demobilization phase of idling can be assumed to be positively correlated with time.This information is based on the distribution of each customer point in time and the assumptions Eq.3.2. Eq.3.8 can thus be used to illustrate the cost of refrigeration during the vehicle transport process:

$$C_3 = \sum_{k=1}^{K}\sum_{i=0}^{N}\sum_{j=0}^{N}\left(at_{ij}^{k}x_{ijk} + bT_iy_{ik}\right) \tag{3.8}$$

T_i is the vehicle's servicing time from customer point .

(4) Price of fresh product cargo loss during vehicle distribution (C_4). Cargo loss is typically associated with the fresh products themselves, as well as distribution collisions, distribution timing, distribution methods, etc. Cold chain distribution, on the other hand, only takes into account cargo loss as a result of the fresh products themselves and the lapse in distribution time because it uses refrigerated vehicles to preserve the products in an atmosphere that is suited for their protection. The two primary factors that contribute to cargo loss are as follows: first, there is cargo loss throughout the distribution process as a result of time and items building up. The other is because of product loading and unloading, which results in environmental changes like temperature and oxygen content variations brought on by the depletion of fresh goods.

This study introduces the freshness decay function of fresh goods:

$$\theta(t) = \theta_0 e^{-\partial t} \tag{3.9}$$

The product's proportion of decay at a given temperature is shown in Eq.3.9. θ_0 represents the freshness of the product, for the product during vehicle transportation, ∂_1 stands for the freshness attenuation coefficient, and ∂_2 represents the freshness attenuation coefficient of the product during vehicle loading and unloading. Due to the loading and unloading process of the carriage door being open to make the temperature in the carriage, oxygen content, and other significant changes in the freshness of the product attenuation rate faster, there is $\partial_2 > \partial_1$. Freshness attenuation coefficient is typically related to the temperature and oxygen content around the goods. In conclusion, the following is an expression of the vehicle distribution process for fresh products during the cost of goods loss:

$$C_{41} = \sum_{k=0}^{K}\sum_{i=1}^{N} y_{ik}Pq_i\left(e^{-\partial_1\left(t_i^k - t_0^k\right)}\right) \tag{3.10}$$

$$C_{42}=\sum_{k=1}^{K}\sum_{i=0}^{N}y_{ik}PQ_{in}\left(1-e^{-\partial_2 T_i}\right) \tag{3.11}$$

$$C_4=\sum_{k=0}^{K}\sum_{i=1}^{N}y_{ik}P\left[q_i\left(e^{-\partial_1\left(t_i^k-t_0^k\right)}\right)+Q_{in}\left(1-e^{-\partial_2 T_i}\right)\right] \tag{3.12}$$

where Q_{in} is the weight of the goods still on board the vehicle as it departs the customer's point i.

4. Create an optimization model for the cold chain distribution path. This study created a mathematical model, represented by Equation Eq.4.1. The concept aims to reduce the overall expenses related to the distribution of cold chain logistics vehicles, including fixed costs, greasing costs, cold storage costs, fresh product loss, and fines for going beyond the time window that was mutually agreed upon.

$$\min Z=C_1+C_2+C_3+C_4+C_5 \tag{4.1}$$

$$\sum_{k=1}^{K}y_{ik}=1,\forall i \tag{4.2}$$

$$\sum_{k=1}^{K}\sum_{j=0}^{N}x_{0jk}=\sum_{k=1}^{K}\sum_{j=0}^{N}x_{j0k} \tag{4.3}$$

$$\sum_{i=0}^{N}x_{ijk}=y_{jk},\forall j,k \tag{4.4}$$

$$\sum_{j=0}^{N}x_{ijk}=y_{ik},\forall i,k \tag{4.5}$$

$$\sum_{i,j\in S\times S}^{j}x_{ijk},S\subseteq\{1,2\cdots N\} \tag{4.6}$$

$$t_j=t_i+T_i+t_{ij},\forall i,j \tag{4.7}$$

Eq.4.2 and Eq.4.7 state that no vehicle's load can exceed the vehicle's maximum load; they also state that a vehicle can only visit each customer point once; they also state that the vehicle must exit the distribution center and return there; they also state that a vehicle may only be permitted to set out and arrive at any one of the customer points once; they are constraints; they seek to eliminate sub-circuits; and they state that the distribution must continue.

5. Sample analyses and solutions. The algorithm's efficacy is tested using Matlab 2018b coding, with the AMD Ryzen 3700x4.2 GHz(16 GB RAM) machine and Win10 as the operating system. The delivery distance of the case is determined by taking the straight line distance from the demand location for the sake of ease in the research.Python is used to scrape the Gaode map and determine the driving time, distance traveled between 66 spots, and real-time road speed using a speed algorithm. The distribution locations of the experiment are sixty-six neighborhoods located 20 km from a cold chain firm in city C. The distribution center is chosen to be the cold chain center in order to ascertain the planar rectangular coordinates of the experiment.

Table 5.1: Examples of medium-scale real data.

No.	Coordinate	Demand	Household Service Time Window	No.	Coordinate	Demand	Household Service Time Window
1	(185.59,89.67)	/	0	18	(140.85,120.80)	2	[3/6,1+30/60]
2	(155.78,78.99)	2.2	[5/6,1+4/6]	19	(137.47,74.87)	1.9	[1/6,1+50/60]
3	(163.22,93.01)	1.8	[15/60,1+2/6]	20	(242.38,100.88)	2.2	[2/6,1+1/6]
4	(153.44,96.65)	1.8	[2/6,50/60]	21	(144.17,140.32)	2.0	[20/60,1+2/6]
5	(142.75,59.32)	2.1	[4/6,1+20/60]	22	(210.55,125.33)	2.4	[5/6,1+20/60]
6	(134.20,62.83)	2.3	[30/60,1+10/60]	23	(138.67,123.44)	2.0	[1+5/6,3]
7	(158.99,157.24)	2.2	[1+2/6,2+3/6]	24	(158.30,98.10)	1.8	[2/6,1+2/6]
8	(141.76,123.11)	2.5	[3/6,1+4/6]	25	(188.95,103.99)	1.7	[3/6,50/60]
9	(130.39,119.72)	2.3	[1+2/6,3]	26	(124.55,78.22)	1.5	[2/6,2+1/6]
10	(189.05,139.33)	1.7	[20/60,1+30/60]	27	(133.47,111.58)	1.9	[1,3]
11	(204.52,72.15)	1.9	[1,1+2/6]	28	(192.33,98.67)	2.2	[2/6,40/60]
12	(136.33,69.55)	1.9	[4/6,2+2/6]	29	(159.84,112.33)	2.4	[1+2/6,2+50/60]
13	(144.88,136.13)	2.8	[1+5/6,3+1/6]	30	(172.68,98.38)	1.8	[4/6,1]
14	(223.17,141.79)	1.6	[40/60,1+4/6]	31	(203.12,115.72)	1.7	[1/6,1+2/6]
15	(216.75,74.92)	1.7	[5/6,50/60]	32	(186.40,100.19)	1.8	[2/6,1+55/60]
16	(152.03,96.87)	2.0	[2/6,1+20/60]	33	(122.45,149.99)	2.5	[1+10/60,3+1/6]
17	(129.82,54.55)	2.5	[1+2/6,2+2/6]	34	(170.87,100.30)	2.2	[40/25/60]

Table 5.2: Initialized pheromone comparison results.

λ	Average distribution cost	Average number of convergences
$\lambda = 1$	1359.8	34.7
$\lambda = 1.5$	1248.6	23.4
$\lambda = 1.8$	1209.5	21.1

The novel knowledge-based ACO algorithm was evaluated and processed on 66 distribution sites to verify its efficacy under varying data quantities.

Table 5.1 displays 34 examples of the demand points, which are set to be l_i (i= 2, 3,...,n) and the distribution center, which is set to be l_1. Based on the obtained real examples, the 33 demand points in medium scale are randomly re-generated according to the probability P by adopting the random offset principle, which produces 30 sets of simulated data for the upcoming analysis and comparison experiments. This process aims to confirm the efficacy of the constructed algorithm and demonstrate its applicability in various scenarios.

5.1. Results analysis. The model is solved using the sample data and the algorithm presented in this work. The model is solved using MATLAB programming; the ideal distribution strategy, which has been executed ten times on a personal computer, is displayed in Fig. 5.1.The number of ants is set to 10,$\alpha = 1, \beta = 3, \rho = 0.5, N_{\max} = 5, \lambda = 1.8, \eta = 0.99$; the total amount of pheromone is 100.

Fig. 5.1 illustrates how effective the study's model algorithm is. Currently, the distribution strategy consists of three vehicles leaving the distribution center: the first vehicle serves the 8th, 13th, 11th, 10th, 12th, 6th, and 15th customer points in that order; the second vehicle serves the 1st, 4th, 5th, 2nd, and 9th customer points; and the third vehicle serves the 3rd, 7th, and 14th customer points. Afterwards, the three vehicles return to the distribution center. Using this technique, the total cost of distribution comes to$1,159.

5.2. Examining the differences between the ACO algorithm. After executing the algorithm ten times at random with the aforementioned parameters, the ACO algorithm with initialized pheromone is compared with the ACO method with uninitialized pheromone. Table 5.2 presents the comparative findings.

Table 5.2 illustrates how the ant colony algorithm's number of convergences is decreased and its speed of convergence is increased when it is initialized with pheromones. Additionally, by combining the A*and ant

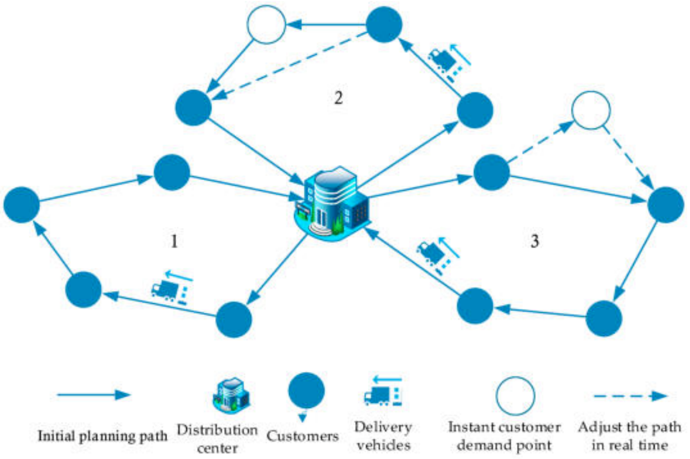


Fig. 5.1: Roadmap for vehicle distribution.

Table 5.3: Results of the experimental comparison.

Experimental	Algorithm	Total Cost	Experimental	Algorithm	Total Cost
1	A	1360	4	A	1360
1	B	1288	4	B	1255
1	C	1217	4	C	1198
2	A	1360	5	A	1360
2	B	1303	5	B	1303
2	C	1240	5	C	1240
3	A	1360	6	A	1360
3	B	1152	6	B	1217
3	C	1130	6	C	1183

Table 5.4: Corresponding optimal distribution techniques for every algorithm.

Algorithm	Minimum Distribution Costs	Distribution Strategy
A	1360	0-6-10-7-12-0
B	1155	0-9-12-13-11-12-6-15-0
C	1130	0-9-12-13-11-12-6-15-0

colony algorithms, the operation's outcomes are optimized, the enterprise's revenue is increased, its distribution costs are decreased, and the enterprise's growth is promoted.

5.3. Comparison of algorithms. Six randomized trials with a total pheromone count of 100, an iteration count of 100, and an initialised pheromone multiplier of 1.8 were carried out for each instance in order to confirm the efficacy of the algorithm, $\eta = 0.99, \alpha = 1, \beta = 3, \rho = 0.4, q_0 = 0.6$, the number of iterations is 100, and the initialization pheromone multiplier is 1.8. Table 5.3- Table 5.5 present the optimal results. Each algorithm's minimum distribution cost and matching distribution method are listed in Table 5.4, and each algorithm's cost composition is provided in Table 5.5 under the minimum distribution cost.

The hybrid ACO algorithm reported in this work outperforms both the A* algorithm and the basic ACO method when energy savings and emission reduction are taken into consideration, according to the experimental results.When compared to the A* algorithm and the basic ACO method, the overall distribution cost is decreased by roughly 15.9% and 8%, respectively. Based on the optimal results' cost components, it is evident

Table 5.5: Algorithm cost components for each algorithm executing the best distribution plan.

Algorithm	Distribution Costs	Fixed Costs	Green Costs	Cargo Damage Costs	Cooling Costs	Penalty Costs
A *Algorithm	1360	445	228	442	38	197
Basic Ant Colony Algorithm	1155	445	193	252	32	222
Hybrid Ant Colony Algorithm	1130	445	188	255	35	209

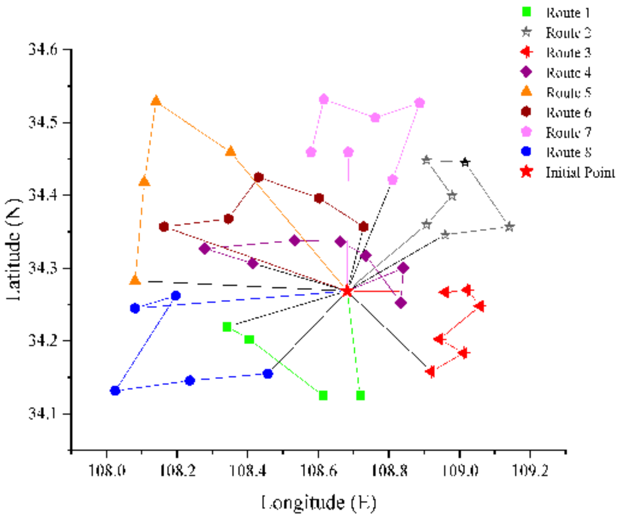


Fig. 5.2: Optimal route map.

that the hybrid ACO algorithm outperforms the original A* algorithm in terms of cost and enterprise operation. This is advantageous for the enterprise's growth as it leverages the ACO algorithm's parallelism and positive feedback to optimize resultsFurthermore, because the cost of green energy and cargo damage has been significantly reduced, it saves resources, safeguards the environment, and maintains product quality all at once. In contrast to the basic ACO algorithm, the hybrid algorithm yields lower green costs and penalty costs, saving resources, protecting the environment, and adhering to the concepts of sustainable development and green logistics. While the overall cost reduction is not as evident, the hybrid algorithm also results in lower carbon pollution and reduced use of environmental resources. Furthermore, the reduction of penalty costs associated with missing delivery windows increases the rate of on-time delivery, which enhances customer satisfaction, the final optimal different routes are shown in Fig. 5.2.

6. Conclusion. Green logistics has emerged as the trend for the future growth of the logistics sector as a result of the ongoing promotion of the idea of sustainable development. This paper presents an analysis of energy saving and emission reduction from the perspective of cost factors to be taken into account in the model. Under the constraints of time window, customer demand, and vehicle loading, a path optimization model is established with the minimum total cost of fixed cost, green cost, refrigeration cost, cost of cargo loss of fresh products, and penalty cost of violating the time window agreed upon with the customer. A hybrid ACO solution model is developed by integrating A^* algorithm to initialize the pheromone of ACO algorithm and reduce the ACO algorithm's convergence time, aiming at the problem of delayed convergence caused by blind search owing to inadequate pheromone at the beginning of ACO algorithm. Examples are used to simulate the algorithm and compare the algorithms in order to confirm the efficacy of both the model and the algorithm. The results demonstrate the effectiveness of both the model and the algorithm and can offer methodological support for the advancement of enterprise search and the application of the concept of green logistics.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

REFERENCES

[1] Tao, N., Yumeng, H., Meng, F.,*Research on cold chain logistics optimization model considering low-carbon emissions.* International Journal of Low-Carbon Technologies, 18,(2023), 354-366.

[2] Li, Y., Xu, X., Guo, X., Liu, X.,*Improved Genetic Algorithm Based Cold Chain Logistics Path Planning with Time Window.* Industrial Engineering and Innovation Management, 6(4),(2023) 44-52.

[3] Wu, L. J., Chen, Z. G., Chen, C. H., Li, Y., Jeon, S. W., Zhang, J., Zhan, Z. H.,*Real environment-aware multisource data-associated cold chain logistics scheduling: A multiple population-based multiobjective ant colony system approach.* IEEE Transactions on Intelligent Transportation Systems, 23(12),(2022), 23613-23627.

[4] Wu, L. J., Shi, L., Zhan, Z. H., Lai, K. K., Zhang, J.,*A buffer-based ant colony system approach for dynamic cold chain logistics scheduling.* IEEE Transactions on Emerging Topics in Computational Intelligence, 6(6),(2022),1438-1452.

[5] Fang, C., Gu, X., Cheng, S., Wu, D.,*Research on long-distance cold chain logistics route optimization considering transport vibration and refrigerant carbon emission.* Procedia Computer Science, 214,(2022), 1262-1269.

[6] Feng, Q., Zhao, G., Li, W., Shi, X.,*Distribution Path Optimization of Fresh Products in Cold Storage Considering Green Costs.* Buildings, 13(9), (2023),2325.

[7] Zhang, T., Xie, W., Wei, M., Xie, X.,*Multi-objective sustainable supply chain network optimization based on chaotic particle—Ant colony algorithm.* Plos one, 18(7), (2023),e0278814.

[8] Ma, Z., Zhang, J., Wang, H., Gao, S.,*Optimization of Sustainable Bi-Objective Cold-Chain Logistics Route Considering Carbon Emissions and Customers' Immediate Demands in China.* Sustainability, 15(7), 5946.

[9] Jia, X.,*Research on the Optimization of Cold Chain Logistics Distribution Path of Agricultural Products E-Commerce in Urban Ecosystem from the Perspective of Carbon Neutrality.* Frontiers in Ecology and Evolution, 10,(2022), 966111.

[10] Zhou, L., Li, Q., Li, F., Jin, C.,*Research on Green Technology Path of Cold-Chain Distribution of Fresh Products Based on Sustainable Development Goals.* Sustainability, 14(24), 16659.

[11] Zhang, G., Dai, L., Yin, X., Leng, L., Chen, H. ,*Optimization of multipath cold-chain logistics network.* Soft Computing, 27(23), 18041-18059.

[12] Ning, T., Gou, T., Liu, X.,*Simulation on cold chain distribution path of fresh agricultural products under low-carbon constraints.* Journal of System Simulation, 34(4),(2022), 797-805.

[13] Abualigah, L., Hanandeh, E. S., Zitar, R. A., Thanh, C. L., Khatir, S., Gandomi, A. H.,*Revolutionizing sustainable supply chain management: A review of metaheuristics.* Engineering Applications of Artificial Intelligence, 126,(2023), 106839.

[14] Cui, H., Qiu, J., Cao, J., Guo, M., Chen, X., Gorbachev, S.,*Route optimization in township logistics distribution considering customer satisfaction based on adaptive genetic algorithm.* Mathematics and Computers in Simulation, 204,(2023) 28-42.

[15] Chunjiong Zhang, Byeong-hee Roh, and Gaoyang Shan. , *Poster: Dynamic Clustered Federated Framework for Multi-domain Network Anomaly Detection*, In Companion of the 19th International Conference on emerging Networking EXperiments and Technologies CoNEXT . Association for Computing Machinery, New York, NY, USA, (2023), 71–72. https://doi.org/10.1145/3624354.3630086.

[16] Qu, L., Li, H.,*Analysis of distribution path optimization algorithm based on big data technology.* Journal of King Saud University-Science, 34(5),(2022), 102019.

[17] Gao, Y., Wu, H., Wang, W.,*A hybrid ant colony optimization with fireworks algorithm to solve capacitated vehicle routing problem.* Applied Intelligence, 53(6),(2023), 7326-7342

[18] Xu, B., Sun, J., Zhang, Z., Gu, R.,*Research on Cold Chain Logistics Transportation Scheme under Complex Conditional Constraints.* Sustainability, 15(10),(2023), 8431.

[19] Li, X., Xie, Q., Zhu, Q., Ren, K., Sun, J.,*Knowledge graph-based recommendation method for cold chain logistics.* Expert Systems with Applications, 227,(2023), 120230.

[20] Jingchun Zhou, Boshen Li, Dehuan Zhang, Jieyu Yuan, Weishi Zhang, Zhanchuan Cai. ,*"UGIF-Net: An Efficient Fully Guided Information Flow Network for Underwater Image Enhancement,"* IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-17, 2023, Art no. 4206117, doi: 10.1109/TGRS.2023.3293912.

[21] Xu, B., Sun, J., Zhang, Z., & Gu, R.*Research on Cold Chain Logistics Transportation Scheme under Complex Conditional Constraints.* Sustainability, 15(10),(2023), 8431

[22] Li, X., Xie, Q., Zhu, Q., Ren, K., & Sun, J. *Knowledge graph-based recommendation method for cold chain logistics.* Expert Systems with Applications, 227,(2023), 120230

[23] Chen, Y., & Qu, X.*Research on Distribution Route Optimization of Logistics Company based on VSP Model.* International Core Journal of Engineering, 9(1),(2023), 23-27.

[24] Chen, J., Kang, H., & Wang, H.*A Product-Design-Change-Based Recovery Control Algorithm for Supply Chain Disruption Problem.* Electronics, 12(12),(2023), 2552.

[25] Zhang, Y.*A low-carbon distribution route selection of supply chain logistics based on internet of things.* International Journal of Internet Manufacturing and Services, 8(3),(2022), 279-294.

[26] Ibrahim, M. F., Mustofa, M. I., Meilanitasari, P., & Wijaya, S. U.*An Improved Ant Colony Optimization Algorithm for Vehicle Routing Problem with Time Windows.* Jurnal Teknik Industri, 23(2),(2022), 105-120.

[27] Wu, C., Xiao, Y., Zhu, X., & Xiao, G. *Study on Multi-Objective Optimization of Logistics Distribution Paths in Smart Manufacturing Workshops Based on Time Tolerance and Low Carbon Emissions.* Processes, 11(6),(2023) 1730.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Apr 1, 2024
Accepted: May 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 180–190, DOI 10.12694/scpe.v26i1.3785

FINE-GRAINED TRUSTED CONTROL METHODS FOR IOT BOUNDARY ACCESS

JIE WANG[*], CHANG LIU[†], GUOWEI ZHU[‡], XIAOJUN LIU[§] AND BIBO XIAO[¶]

Abstract. In order to solve the problems of coarse-grained access policies, weak auditability, lack of access process control, and excessive privileges exhibited by existing IoT access control technologies, the author proposes a fine-grained trusted control method for IoT boundary access. The author elaborated on the CcBAC model framework and formalized its definition; At the same time, specific descriptions of the functions in the model were provided, and the access control process of this model in general application scenarios was presented; After rigorous testing, it was found that as the number of requests increased, there was a slight uptick in the average time it took for the function to process them, but beyond a certain point, this time plateaued and even began to decrease gradually. Meanwhile, the system's throughput increased steadily with more requests until it reached a stable level, showing no significant drop even with additional clients. The proposed CcBAC access control model showcased remarkable performance in handling large-scale requests while ensuring fine-grained, autonomous authorization, security, and auditability. It effectively achieved consensus in distributed systems and maintained data consistency. In conclusion, this model empowers resource owners with full control over their resources' access, while also accounting for the detailed and traceable nature of access control.

Key words: Blockchain, Access control, Internet of Things, Cryptocurrency, Trusted Execution Environment

1. Introduction. With the development and maturity of Internet of Things technology, it has gradually evolved from an initial concept and penetrated into various fields, such as intelligent transportation, medical care, environmental monitoring, and other multi industry fields [1]. Billions of IoT devices have begun to be widely used in people's social lives, generating exponential growth in IoT data. The era of data has arrived. The Internet of Things data is the core driving force behind the Internet of Things, containing enormous value, and the most important way to unleash the value of data is to make it flow. In order to accelerate the process of data sharing and circulation, and better serve various fields such as education, business, and infrastructure, in recent years, the country has elevated the construction and development of data to a national strategy and actively transitioned towards a digital economy [2]. In October 2020, the Central Committee of the Communist Party of China unveiled the "14th Five-Year Plan for National Economic and Social Development." This plan emphasized the imperative of advancing the utilization of data and information resources, with a specific goal of fostering open sharing of fundamental information. In March 2021, the State Council released the Government Work Report, which mentioned the need to improve data sharing coordination capabilities and improve data sharing coordination mechanisms [3]. Driven by policies, there is an increasing amount of research on data sharing in the Internet of Things, and a number of data sharing application cases have emerged. However, due to the fact that most solutions adopt a centralized data management model, the degree of sharing is low, and data cannot be monetized, so the value of data cannot be released. The reason behind this is that there are currently many problems exposed in the process of data sharing transactions, lacking a reliable, open and transparent data sharing atmosphere [4]. Blockchain, relying on features such as decentralization, difficulty in tampering, and traceability of transactions, provides the possibility of building a reliable and transparent IoT data sharing environment.

2. Literature Review. As a new decentralized, trustworthy, and tamper proof technology, blockchain can provide a trusted implementation solution for trust and data security issues in the field of data sharing.

[*] State Grid Hubei Electric Power Research Institute, Hubei Wuhan, 430077, China (Corresponding author, JieWang59@163.com)

[†] State Grid Hubei Electric Power Research Institute, Hubei Wuhan, 430077, China (ChangLiu61@126.com)

[‡] State Grid Hubei Electric Power Co., Ltd, Hubei Wuhan, 430077, China (GuoweiZhu7@163.com)

[§] State Grid Yichang Power Supply Company, Hubei Yichang, 443000, China (XiaojunLiu8@126.com)

[¶] State Grid Yichang Power Supply Company, Hubei Yichang, 443000, China (BiboXiao6@163.com)

Therefore, the combination of blockchain technology and IoT data sharing has become a research hotspot in the academic community [5].

The characteristics of blockchain can provide an auditable platform for data sharing, such as recording the data sharing process on the chain, providing a data foundation for subsequent service evaluation. The combination of blockchain and the Internet of Things has received considerable academic research, and in the field of data sharing, blockchain mainly focuses on protecting the confidentiality, integrity, and availability of data. Atlam, H. F. et al. introduced an innovative IoT technology access control model centered around risk assessment, catering to real-time data requests from IoT devices while offering dynamic responsiveness. This model leverages IoT environment attributes to gauge security risks linked to access requests, employing factors such as user context, resource sensitivity, action severity, and past risk records to inform its security risk estimation algorithm, pivotal for access control decisions. Moreover, the model integrates smart contracts to enable adaptability, actively monitoring authorized user conduct to swiftly identify any irregular behaviors[6]. Wang et al. introduced an access control approach tailored for laboratory cloud data, integrating Internet of Things (IoT) technology. The method breaks down laboratory cloud database data into minimal attributes, encrypting them and creating a key of minimal granularity that adheres to access tree constraints. By combining IoT and proxy re-encryption technologies, the method maps access structures and attribute sets using hash functions, encrypts symmetric keys via the CP-ABE scheme, and facilitates laboratory cloud scalability based on access control methodologies[7]. Conventional centralized data sharing systems pose risks like single points of failure and overburdened central nodes. To address these challenges and foster a more distributed and collaborative approach, researchers have increasingly turned to blockchain-based solutions for Internet of Things (IoT) environments. However, without predefined policies, legitimate user access may be denied, and data updates on the blockchain may bring high costs to owners. Wang, R. et al. addressed these challenges by integrating the Accountable Subgroup Multiple Signature (ASM) algorithm with Attribute Based Access Control (ABAC) techniques, along with policy smart contracts. This fusion delivers a precise and adaptable solution, offering detailed control over access permissions [8].

Traditional models such as RBAC, ABAC, and CapBAC have limitations in fine-grained control, policy flexibility, and auditability, while blockchain based access control schemes have significant advantages in data immutability and decentralization, despite advancements, there are opportunities to enhance policy articulation and execution speed. To address these, the author introduces a novel access control model leveraging cryptocurrency and trusted execution environments (CcBAC). This model automates policy determination and transparently executes them via blockchain, while ensuring secure off-chain policy execution through trusted execution environments (TEE). The CcBAC model not only solves the application problems of existing methods in the Internet of Things environment, but also provides new ideas and methods for future related research. Through this combination, we aim to provide a more secure, flexible, and auditable access control solution for IoT devices.

3. Research Methods.

3.1. CcBAC Access Control Model. The CcBAC access control model is a universal access control model that uses the cryptocurrency Ccoin to represent access permissions. The purpose of Ccoin is to concretize access capabilities into atomic, trustworthy, and transferable digital assets. The core idea is to establish a unified standard to describe the permission control policies of each resource through deep integration of trusted execution environment and blockchain technology, in order to achieve fine-grained and dynamic management of policies and provide convenience for user access. Using blockchain networks as interaction media, driven by smart contracts, to achieve automated decision-making of access policies, authentic and trustworthy execution, transparent and traceable access processes, and auditable policy execution processes. By introducing trusted access control objects, control over the access process is achieved to prevent excessive privileges; Meanwhile, any user can independently formulate access policies to achieve flexible management and sharing of resources.

In this model, Ccoin is a cryptocurrency used for access control, which is used to verify and authorize user access to resources. Each Ccoin contains access policies and metadata to achieve fine-grained access control. Ccoin plays the following roles in access control security:

1) Fine grained control: Each Ccoin includes access policies that can define specific access permissions and restrictions. Through Ccoin, fine-grained control of resources can be achieved, allowing only authorized

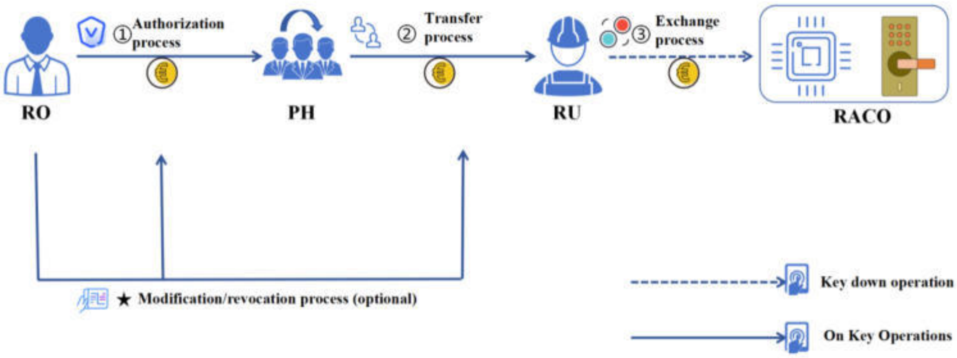


Fig. 3.1: Access Control Model Based on Ccoin

users to perform specific operations [9].

2) Auditability: The operations and resource access activities of Ccoin are recorded in blockchain transactions to ensure auditability. This means that all operations and access activities are public, verifiable, and traceable.
3) Access process control: Through Ccoin, control over the access process can be achieved. Before accessing resources, a verification and authorization process is required to ensure that users meet the access conditions. At the same time, Ccoin can also record activities during the access process for auditing and control.
4) Preventing Witch Attacks: Ccoin can effectively prevent witch attacks through the decentralized nature of blockchain technology.
5) Secure storage: The operations and access activities of Ccoin are recorded in the transactions of the blockchain, which means that Ccoin's data is stored in a distributed network. Compared to traditional centralized storage, distributed storage has higher security and reliability.

The access strategy consists of five key fine-grained elements that determine the access control scheme: In access control terminology, the "4W1H access policy" refers to specifying the "who, what, where, when, and how" of access permissions. Essentially, it outlines the conditions necessary for granting access requests. The "who" identifies the authorized users, the "what" specifies the actions they are permitted to perform, the "where" indicates the locations where access is granted, and the "when" delineates the time frames for access. Finally, the "how" details the precise process that must be adhered to during the access period. The 4W1H access policy is forged into the digital currency Ccoin, and all operations against Ccoin will be securely recorded for review throughout its entire lifecycle. Before the Ccoin is redeemed, resource owners can modify or revoke access policies in the Ccoin to achieve fine-grained control over resource access permissions[10]. This approach exhibits four distinct traits:

1) Sole Authority: Only resource owners possess the ability to create, modify, and revoke Ccoins.
2) Transferability: Holders of Ccoins can freely transfer them to other participants.
3) Conditional Redemption: Ccoin redemption is contingent upon meeting access policy conditions, with strict adherence required for resource access.
4) Immutable Recording: All access actions are securely logged on the blockchain for auditing purposes. The model facilitates interactions between Resource Owners (RO), Resource Requesters (RU), Permission Holders (PH), and Reliable Access Control Objects (RACO), with each entity identified by distinct addresses. Participants possess individual wallets containing Ccoins, representing access permissions. Transactions within this model primarily encompass three steps, as illustrated in Figure 3.1.

The trusted access control object, RACO, serves as a dependable module representing the interests of resource owners. Its responsibilities include verifying the functional integrity of Ccoin, ensuring compliance with fine-grained access conditions, making access decisions, and monitoring the access process [11]. To ensure the trustworthiness and security of RACO, it is housed within a trusted execution environment (TEE). TEE offers a secure processing environment that is resistant to tampering, allowing for secure isolation and storage within the chipset. Its core functionalities include secure startup, runtime isolation, secure storage, scheduling, trusted I/O operations, and remote management. TEE can ensure the authenticity, integrity, and confidentiality of RACO code and runtime status, and resist software attacks and physical tampering.

3.2. Basic Elements and Functions.

Definition 1. CcABC defines the cryptocurrency representing access rights as Ccoin, denoted as c=(tokenId, owner, holder, device, policy, timestamp, isValid).

This approach leverages blockchain technology to ensure secure data storage and atomic data transmission, with consensus among participating nodes maintaining ledger consistency. Ccoin transactions take place on the blockchain, with operations such as createCcoin, transferFrom, updatePolicy, revokeCcoin, and redeemCcoin initiated by sending messages to all blockchain nodes. Ensuring authenticity, the function caller must sign the message using a key. In CcBAC, the message format of the caller's public key infrastructure (pki) is formatted as follows 3.1:

$$msg : [tokenId, op, \{policy\}, \{pk_j\}]_{\sigma_{pk_i}} \tag{3.1}$$

Among them, op represents the operation code of the function, and optional information is enclosed in curly braces. If op=createCcoin or updatePolicy, $\{policy\}$ holds a valid policy; If op=transferFrom, then $\{pk_j\}$ is the new recipient public key. σ_{pk_i} represents the key signature of the function caller pk_j.

System participants include both regular blockchain nodes and users who interact with the blockchain through secure channels. When participants engage in actions such as creating, transmitting, modifying, revoking, or redeeming Ccoins, they send signed messages to the blockchain, which subsequently authenticates the messages to ensure their legitimacy. Before any Ccoin operation, the functionality of the function caller is checked. Ccoins remain on the blockchain until redemption, and each Ccoin operation triggers a new transaction. These transactions include the Ccoin and a script detailing the call message and resulting activities, facilitating review[12].

3.3. Workflow. CcBAC distinguishes itself from other access control models in three main ways. Firstly, it enables fine-grained access control through comprehensive access policies that specify who, what, where, when, and how access is granted. These policies ensure secure access by defining strict procedures to follow, while access permissions are represented as digital assets called Ccoins, stored on the blockchain and managed through secure and atomic operations, akin to cryptocurrency transactions. Secondly, access policy verification occurs within a Trusted Execution Environment (TEE) security zone, providing physical protection for relevant programs and securely collecting environmental evidence to facilitate accurate access decisions and monitor the access process. Lastly, these design elements ensure that CcBAC offers fine-grained and responsible access control with encryption-level security trust, effectively preventing unauthorized access. The workflow of CcBAC can be outlined in the following steps:

Step 1: Send the request Participants send operational requests for Ccoin, including transmission, revocation, update, and redemption.

Step 2&Step 3: Verify that the blockchain verifies the request, including message signature, Ccoin validity, and participant permissions on Ccoin. If it is a redemption request, RACO needs to further verify the environmental data.

Step 4: Access and monitoring. After obtaining access permissions, RACO will access resources and monitor access activities based on policies.

Step 5: Record. Store the accessed records in the blockchain for auditing purposes.

The verification process in CcBAC consists of two main components. Firstly, it verifies identity and transaction authenticity through the blockchain. Secondly, it validates access control policies via the policyCheck function within the authorization process services, which relies on the trusted access control object RACO implemented using Trusted Execution Environment (TEE) technology. Upon receipt of an access request, RACO collects data from the physical environment and securely communicates with the blockchain system to extract access policies from the corresponding Ccoins. To efficiently manage Ccoins within distributed ledgers, the

Table 4.1: Experimental Environment and Configuration

Configuration	Specific information
operating system	Ubuntu 18. 04. 1 GNU/Linux
CPU	8 Intel (R) Core (TM) i7-6700HQ CPU @ 2. 6GHz
network card	Intel Corporation Ethernet Connection (2) I219-LM (rev 31)
Memory	16G Samsung PC4-2400T-UA2
Hard disk	512G SSD Samsung SSD, 1T HDD ST1000DM003-1SB1 CC4
TEE chipset	ARMv8-M TrustZone, LPC55S69-EVK

UnRedeemed Policy Output (URPO) model, modeled after Bitcoin's Unspent Transaction Output (UTXO) model, has been developed. This model organizes multiple Ccoins within each block, meticulously recording their metadata and transactional activities within the TX script fields. Access policies in Ccoin are represented using JSON key-value pairs, enabling flexible policy granularity. CcBAC inherently incorporates auditability and traceability, as every operation and resource access activity is meticulously logged in the TX script field of transactions stored on the blockchain. This information is openly accessible and serves to provide a comprehensive record of system activities, and validated by the entire blockchain network[13,14].

4. Result analysis.

4.1. Experimental analysis. The CcBAC system, as proposed by the author, comprises two key components: a blockchain-based distributed ledger and a trusted access control object leveraging TEE chipsets. The blockchain serves as a secure platform for managing Ccoins, facilitating Ccoin operations through transactions, and recording all activities for auditability purposes. Meanwhile, the trusted access control object integrates blockchain clients and sensor drivers within its secure zones, enabling it to gather environmental data and make reliable access control decisions.

In the second component of the system, the experiment employed the LPC55S69-EVK microcontroller in conjunction with sensors including a GPS receiver, temperature sensor, and camera, all safeguarded by ARMv8-M TrustZone. ARMv8-M TrustZone, renowned for its energy-efficient design, consists of a secure zone, a non-secure callable interface, and a non-secure zone. Unlike traditional architectures that isolate non-secure zones from secure resources, TrustZone allows security codes in secure zones to have elevated permissions, enabling access to resources in both secure and non-secure zones. To ensure the secure collection of environmental data essential for access policies, the experiment incorporated a sensor driver within the secure zone of the TEE. This driver facilitates direct communication with sensor hardware. Additionally, a lightweight blockchain client was integrated into the security zone of the TEE, enabling secure communication with the blockchain through SSL/TLS protocols[15]. The experimental setup and configurations are elaborated in detail in Table 4.1.

4.1.1. Adaptability analysis. To gauge the versatility of the author's prototype system across various mainstream platforms, the experiment deployed the prototype on three separate platforms: Golang, the Ethereum main network, and the Ethereum-based consortium chain Quorum. Each function within the model underwent independent testing 50 times, and the average runtime for each function on different platforms was recorded. The performance testing of various functions within the Golang prototype revealed relatively uniform time distribution and stable performance. Notably, functions such as createCcoin, transferFrom, updatePolicy, and revokeCcoin demonstrated relatively brief runtimes, spanning from approximately 30 to 80 milliseconds. Conversely, the redeemCcoin function demanded more time due to its involvement in policyCheck. This process entails the sampling and analysis of sensor readings by the access control object of CcBAC, as well as the execution of necessary actions and transmission of data back to the blockchain, thereby resulting in a longer runtime.

Figure 4.1 illustrates the total time duration for each Ccoin operation function across the three prototype systems. Utilizing a logarithmic scale, the figure effectively visualizes the considerable difference in confirmation times between Go Coin and Ethereum Ccoin on various platforms. Upon closer examination, it becomes appar-

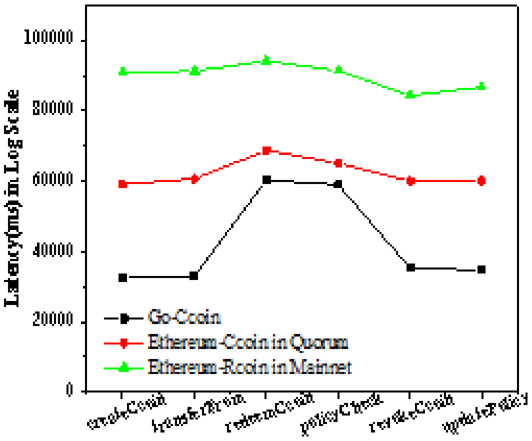


Fig. 4.1: Running Time of Each Function under Three Prototypes

ent that native Go Coins generally require only 40-60 milliseconds to confirm each transaction. In comparison, Ethereum Ccoins within the Quorum consortium chain require approximately 1 second for confirmation, while on the Ethereum main network, this confirmation time extends to approximately 30 to 50 seconds. These findings underscore the good adaptability of the CcBAC model proposed by the author across different platforms.

4.1.2. Performance analysis. To assess the performance of the CcBAC system, the experiment utilized the Go Coin prototype and conducted simulations with multi-threaded clients to simulate concurrent access. Two sets of comparative experiments were designed for this purpose. In the first set, the processing time of each function was measured for varying numbers of concurrent requests, ranging from 50 to 1000 virtual clients. The findings, as depicted in Figure 3-4, reveal a consistent pattern. Figure 3 demonstrates that as the number of requests escalates from 50 to 500, the total runtime of each function steadily rises, albeit at a diminishing rate beyond 500 requests. Meanwhile, Figure 4 illustrates a slight increase in the average time cost of functions as the number of requests climbs from 50 to 200. However, with additional requests, there's a subsequent decrease in average time cost, ultimately stabilizing over time[16]. These results suggest that the system's throughput grows as the number of requests increases. Once a certain threshold is reached, throughput stabilizes, with no significant decline observed even as the number of clients continues to rise.

4.2. Security Analysis. In real-world scenarios, access control systems are vulnerable to two main types of attacks: access permission forgery and access policy violations. To model the competitive dynamics between honest nodes and attacking nodes, the author employs a binomial stochastic process. This process determines that an attack is successful only if the block length created by the attacking node surpasses the length created by the honest node. If the probability of the attacker ultimately winning the game is p, then both the attacker's victory and failure need to meet certain conditions. Assuming that the probability of the attacker succeeding in each operation is q and the probability of the honest node succeeding in each operation is p, the probability of the attacker winning p can be calculated using the Gambler's Ruin Problem (GRP) formula. Since k may be any integer greater than or equal to 0, the probability of k taking a value follows a Poisson distribution, and the probability of k occurring is equation 4.1:

$$p_k = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.1}$$

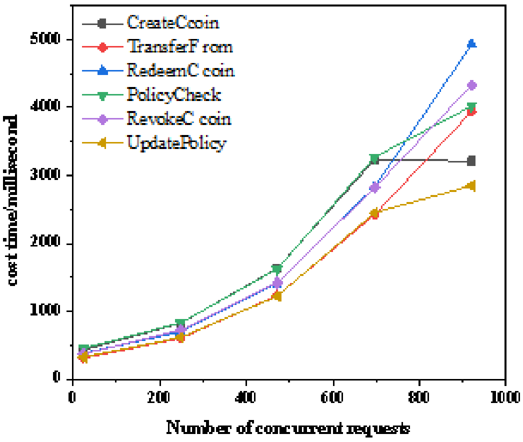


Fig. 4.2: Time cost of functions under different concurrent request quantities

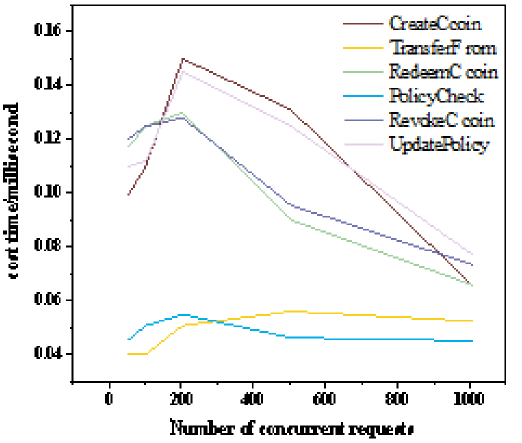


Fig. 4.3: Trend of Average Time Cost of Functions under Different Concurrent Requests

Among them, λ is the mean of k, and the ratio relation equation 4.2 below is satisfied:

$$\lambda = z \cdot \frac{q}{p} \tag{4.2}$$

When the tampering chain extends k blocks, the probability of catching up with the honest chain is equation 4.3:

$$q_z = f(x) = \begin{cases} 1, p \leqslant q \\ (\frac{q}{p})^{z-k}, p > q \end{cases} \tag{4.3}$$

Compute the probability of all values from k to positive infinity, then add them up, and finally calculate

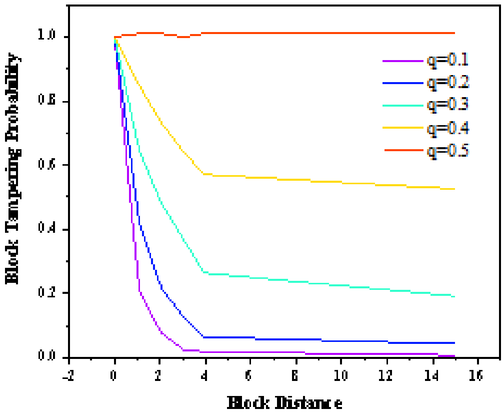


Fig. 4.4: Probability of successful attack

the probability formula 4.4 for attackers to successfully tamper with block data:

$$P = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \cdot \begin{cases} (\frac{q}{p})^{z-k}, k \leqslant z \\ 1, k > z \end{cases} \tag{4.4}$$

To avoid infinite sequence summation when calculating the final result, further transform it into equation 4.5:

$$P = 1 - \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \cdot (1 - (\frac{q}{p})^{(z-k)}) \tag{4.5}$$

The author conducted simulation experiments using MATLAB to analyze the relationship between the probability of block tampering (P) and the number of blocks (n), as illustrated in Figure 4.4. As the number of blocks in the blockchain increases, the probability of successful attacks diminishes rapidly. It becomes evident that attackers require control of more than 50% of the nodes in the blockchain to successfully tamper with the next block. This implies that the blockchain-based IoT device access control mechanism is highly effective in thwarting access permission forgery attacks. Given the vast number of devices in the IoT ecosystem, each capable of serving as a light node in the blockchain, the security of access control is bolstered. Consequently, the author's approach to blockchain-based IoT device access control effectively mitigates forgery attacks on access permissions, aligning with the stringent security requirements of the IoT ecosystem for device access control [17,18].

In addressing the second type of attack, the author considers access policy violations stemming from weak security in IoT devices themselves. Many IoT devices lack robust security measures, making them vulnerable to breaches in code confidentiality and integrity. Attackers exploit these vulnerabilities in lightweight IoT devices to undermine access control security. To mitigate such risks, the author's CcABC model harnesses Trusted Execution Environment (TEE) to deploy reliable access control objects. This strategy establishes a secure enclave for executing code and safeguarding data, guaranteeing privacy and trustworthiness. TEE interfaces with the physical world to establish secure links with the blockchain, extending trust from on-chain to off-chain activities. By securely retrieving access policies from Ccoins stored in the blockchain, the model makes access decisions based on predefined access conditions and rigorously monitors the access process. This ensures that resource visitors adhere strictly to access policies set by the resource owner, effectively thwarting access policy violations. Another significant factor contributing to policy violation attacks is the coarse-grained nature of the access control model itself. Unlike the author's PBAC access control model, which ensures sufficient granularity, other models primarily focus on users rather than resources. For instance, conventional models may inquire, ”Which users are present, and what resources can they access?” This approach typically involves a subject (the entity seeking access), permissions (the actions permitted), and roles (sets of permissions assigned to subjects). In contrast, the author's model delves into, ”What types of users exist, and what actions can they perform within the environment?” Here, the control comprises a principal (the entity seeking access), permissions (the actions permitted), and roles (sets of permissions assigned to the principal). By defining access control policies based on who, what, where, when, and how, the author's model achieves finer granularity, effectively mitigating the risk of access policy violations.

4.3. Comparative analysis. The author evaluates the proposed access control model from two perspectives. Firstly, by comparing it with traditional models, the advantages and disadvantages of the CcBAC model are assessed. It's evident that this model offers distinct advantages in handling the complexities of the modern Internet of Things, particularly in scenarios involving massive scale, dynamic environments, and distributed systems.

On the other hand, compared with existing research models, analyze the advantages of the CcBAC model. The access control model described by the author has the following advantages:

Fine-grained CcBAC access control meticulously specifies who, what, where, when, and how—five crucial elements that shape the precision of access control schemes. This approach effectively answers the question of who has permission to perform specific actions, when, and where.

In terms of security, the model conducts all operations related to Ccoin on the blockchain, ensuring that every message is authenticated via a secure signature mechanism. This approach supports trusted storage and guarantees atomic state transitions. Additionally, a dependable access control object, RACO, is responsible for securely retrieving access policies from Ccoins stored on the blockchain. It then makes access decisions based on predefined access conditions and diligently monitors the entire access process.

Convenient access for users. Accessing a specific resource is straightforward—all you need is to redeem the corresponding resource using Ccoin. Additionally, audit procedures enable flexible transfer of Ccoin from one holder to another, facilitating dynamic transfer of access permissions.

Autonomous authorization. The model ensures that access to resources is solely determined and granted through Ccoin by the resource owner whenever necessary, without any involvement or intervention from third parties accessing the server.

Auditability. The author's CcBAC model has native auditability, as all operations and resource access activities through Ccoin are recorded in the TX script fields of transactions stored in the blockchain, and are publicly and validated by the entire blockchain system.

Access Process Control. RACO serves as a dependable access control entity, embodying the responsibilities of resource owners. Apart from validating the functionality of Ccoin and ensuring compliance with fine-grained access conditions, RACO actively monitors the entire access process. Through integration with TEE, it guarantees the secure recording of all activities occurring during access, providing comprehensive oversight and security assurance.

5. Conclusion. In practical applications, the author's access control model can be widely applied to various devices in the IoT ecosystem. For example, in industrial control systems, the CcBAC model utilizes the decentralized nature of blockchain technology to create unique digital identities for each device, enabling effective authentication and permission management even in the case of a large number of devices. Through smart contracts, permission allocation and revocation between devices can be automated, reducing the complexity and error rate of manual configuration. In industrial automation, the roles of operators and equipment may dynamically change according to production needs. The CcBAC model allows for quick adjustment of access policies through smart contracts to adapt to these changes, ensuring that only authorized operators can access the corresponding devices and data. The CcBAC model utilizes blockchain technology to achieve unified access control policies for cross regional devices, ensuring consistency and security in distributed environments. The immutability of blockchain ensures the integrity of access records, and any unauthorized access attempts will be recorded and traceable. Industrial automation has strict requirements for real-time response, and the CcBAC model reduces transaction confirmation time and improves system response speed through the fast consensus mechanism of blockchain. The Trusted Execution Environment (TEE) provides a secure environment for performing sensitive operations, such as verifying and executing access policies. This ensures that access control policies can be securely executed even in environments where devices may be vulnerable to attacks.

In summary, the blockchain based IoT device access control model can effectively prevent access permission forgery attacks and meet the security requirements of the IoT ecosystem for device access control. When implementing this model, enterprises need to consider the following suggestions for action:

(1) Enhance IoT Device Security: Implementing a trusted execution environment safeguards the confidentiality and integrity of IoT devices, shielding them from potential exploitation by malicious actors.

(2) Flexible formulation of access policies: Based on actual needs, develop fine-grained access policies to ensure that resource visitors strictly adhere to access policies and prevent the occurrence of access policy violations.

(3) Implement unified access control policy standards: Through blockchain storage policies, it facilitates information sharing and supports unified access control policy standards for all parties, improving system scalability and interoperability.

Future direction. Focus on implementing Organizational Based Access Control (CcBAC) in specific application scenarios to address specific challenges in this field. This includes exploring how to design and deploy access control policies targeting specific business needs and security threats to ensure the security and availability of the system. Conduct in-depth research on specific needs in different industries or fields, and develop corresponding solutions to meet customer needs and improve overall efficiency and security of the system.

Acknowledgement. This work was supported by 2022 science and technology project of State Grid Hubei Electric Power Co., Ltd, ”research and demonstrational application of key technologies of industrial control security proactive defense architecture for novel power systems”(No.52153222001D).

REFERENCES

[1] Fan, Y., Liu, S., Tan, G., & Qiao, F. (2020). Fine-grained access control based on trusted execution environment. Future Generation Computer Systems, 109, 551-561.

[2] Jiang, W., Li, E., Zhou, W., Yang, Y., & Luo, T. (2023). IoT access control model based on blockchain and trusted execution environment. Processes, 11(3), 723.

[3] Lee, S., Jo, H. J., Choi, W., Kim, H., Park, J. H., & Lee, D. H. (2020). Fine-grained access control-enabled logging method on ARM TrustZone. IEEE Access, 8, 81348-81364.

[4] Zhou, Q., Elbadry, M., Ye, F., & Yang, Y. (2020). Towards fine-grained access control in enterprise-scale Internet-of-Things. IEEE Transactions on Mobile Computing, 20(8), 2701-2714.

[5] Chattaraj, D., Bera, B., Das, A. K., Rodrigues, J. J., & Park, Y. (2021). Designing fine-grained access control for software-defined networks using private blockchain. IEEE Internet of Things Journal, 9(2), 1542-1559.

[6] Atlam, H. F. , Alenezi, A. , Walters, R. J. , Wills, G. B. , & Daniel, J. . Developing an Adaptive Risk-Based Access Control Model for the Internet of Things. IEEE International Conference on Green Computing and Communications;IEEE International Conference on Cyber, Physical and Social Computing;IEEE International Conference on Internet of Things;IEEE International Conference on Smart Data, 15(6), 3485-3498.

[7] Wang, L. , & Yu, Y. . (2023). Access control method of laboratory cloud data based on internet of things technology. Int. J. Auton. Adapt. Commun. Syst., 16, 31-47.

[8] Wang, R. , Wang, X. , Yang, W. , Yuan, S. , & Guan, Z. . (2022). Achieving fine-grained and flexible access control on blockchain-based data sharing for the internet of things. China Communications (English version), 19(6), 22-34.

[9] Liu, X., Wang, H., Zhang, B., & Zhang, B. (2022). An efficient fine-grained data access control system with a bounded service number. Information Sciences, 584, 536-563.

[10] Yu, D., Hsu, R. H., Lee, J., & Lee, S. (2022). EC-SVC: Secure can bus in-vehicle communications with fine-grained access control based on edge computing. IEEE Transactions on Information Forensics and Security, 17, 1388-1403.

[11] Lee, U., & Park, C. (2020). SofTEE: Software-based trusted execution environment for user applications. IEEE access, 8, 121874-121888.

[12] Li, H., Pei, L., Liao, D., Chen, S., Zhang, M., & Xu, D. (2020). FADB: A fine-grained access control scheme for VANET data based on blockchain. IEEE Access, 8, 85190-85203.

[13] Pareek, G., & Purushothama, B. R. (2020). Proxy re-encryption for fine-grained access control: Its applicability, security under stronger notions and performance. Journal of Information Security and Applications, 54, 102543.

[14] Cao, Q., Li, Y., Wu, Z., Miao, Y., & Liu, J. (2020). Privacy-preserving conjunctive keyword search on encrypted data with enhanced fine-grained access control. World Wide Web, 23, 959-989.

[15] Yin, H., Qin, Z., Zhang, J., Deng, H., Li, F., & Li, K. (2020). A fine-grained authorized keyword secure search scheme with efficient search permission update in cloud computing. Journal of Parallel and Distributed Computing, 135, 56-69.

[16] Zhang, W., Liu, S., & Xia, Z. (2022). A distributed privacy-preserving data aggregation scheme for smart grid with fine-grained access control. Journal of Information Security and Applications, 66, 103118.

[17] Zhang, X., Shi, R. H., Guo, W., Wang, P., & Ke, W. (2023). A dual auditing protocol for fine-grained access control in the edge-cloud-based smart home. Computer Networks, 228, 109735.

[18] Oh, H., Nam, K., Jeon, S., Cho, Y., & Paek, Y. (2021). MeetGo: A trusted execution environment for remote applications on FPGA. IEEE Access, 9, 51313-51324.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Apr 23, 2024
Accepted: Jun 26, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 191–200, DOI 10.12694/scpe.v26i1.3813

A NEURAL COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM BASED ON ATTENTION MECHANISM AND CONTRASTIVE LEARNING

JIANQIAO LIU*

Abstract. The neural collaborative filtering recommendation algorithm is widely used in recommendation systems, which further applies deep learning to recommendation systems. It is a universal framework in the neural collaborative filtering recommendation algorithm, however, it does not consider the impact of different features on recommendation results, nor does it consider the issues of data sparsity and long tail distribution of items. To solve the above problems, this paper proposes a recommendation algorithm based on attention mechanism and contrastive learning, which focuses on more important features through attention mechanism and increases the number of samples to achieve data augmentation through contrastive learning, therefore it improves model performance. The experimental results on two benchmark datasets show that the algorithm proposed in this paper has further improved recommendation performance compared to other benchmark algorithms.

Key words: attention mechanism; contrastive learning; deep learning; neural collaborative filtering; recommendation algorithm

1. Introduction. With the development of information technology, today's society has entered the era of informatization, generating a large amount of data every moment. How to quickly and accurately obtain information that users are interested from these data is a challenge, and recommendation systems can effectively solve this problem. At present, recommendation systems have been widely applied in e-commerce, social networks, news communication, and other fields [1-3]. Recommendation algorithms are an important component of recommendation systems, and their performance directly affects the recommendation effectiveness. Currently, collaborative filtering recommendation algorithms are a widely used recommendation algorithm that recommends items of interest to users by calculating their similarity and item similarity[4]. Among various collaborative filtering recommendation algorithms, matrix factorization is one of the popular ones. Matrix factorization decomposes user and item interaction information into two low dimensional user matrix and item matrix, obtains the potential relationship between users and items, and recommends items of interest to users through the product of the user matrix and item matrix [5]. In recent years, deep learning has also developed rapidly, and neural networks in deep learning have achieved great success in many fields. Many researchers have attempted to apply neural networks to the field of recommendation and have achieved some good results, among which the neural collaborative filtering recommendation algorithm is one of them [6]. The neural collaborative filtering recommendation algorithm is a recommendation algorithm that applies neural networks to matrix factorization. It is a universal recommendation algorithm framework with both linear and nonlinear modeling capabilities, which can learn stronger user and item representation abilities and has good recommendation performance.

In recent years, with the continuous increase in data, it has been difficult to label data, and many data do not have labels. Contrastive learning belongs to self-supervised learning, which can learn from unlabeled data and extract important features from unlabeled data. Contrastive learning has made tremendous achievement in many application fields [7-9]. In the field of recommendation systems, contrastive learning can transform positive samples, add new samples to the sample space, and enhance the representation of data. By using contrastive learning, the recommendation performance of recommendation systems can be effectively improved. The attention mechanism simulates human attention, focusing more on important things and less on unimportant things. In recommendation systems, utilizing attention mechanisms can provide more attention and weight to important users and items, rather than treating all users and items equally [10-12]. The attention mecha-

*School of Foreign Languages, Fuyang Normal University, Fuyang Anhui 236037, China (ianqiaoliu@tom.com)

nism can obtain important information, grasp key parts, better train models, and improve model performance. Therefore, we proposes an improved neural collaborative filtering recommendation algorithm (NCF-AC), which integrates attention mechanism and contrastive learning on the basis of neural collaborative filtering algorithm. This algorithm uses attention mechanism to adjust the weights of different items, concentrate more attention on important items, and use contrastive learning to add new samples for data augmentation. Experiments were conducted on public datasets, and the results showed that the recommendation algorithm proposed in this paper is effective. Compared with other recommendation algorithms, the recommendation algorithm proposed in this paper has better recommendation performance.

The contributions of this paper can be summarized as follows:

1. This paper proposes an improved neural collaborative filtering recommendation algorithm (NCF-AC), which integrates attention mechanism and contrastive learning on the basis of the neural collaborative filtering algorithm. Use attention mechanism to adjust the weights of different items and focus more attention on important items. Use contrastive learning to add new samples for data augmentation.
2. Experiments were conducted on public datasets, and the results showed that the recommendation algorithm proposed in this paper is effective. Compared with other recommendation algorithms, our proposed recommendation algorithm has better recommendation performance.

The rest of this paper is organized as follows. In Section 2, we introduced the neural collaborative filtering model. In Section 3, we provided a detailed introduction to our proposed algorithm. In Section 4, we conducted experiments on two public datasets to validate the effectiveness of our proposed algorithm. Finally, in Section 5, we gave a conclusion of this paper.

2. Neural collaborative filtering. In recent years, neural networks have developed rapidly and achieved great success in image processing, voice recognition, natural language processing, and other fields. Neural collaborative filtering (NCF) is a collaborative filtering recommendation algorithm based on neural network technology, which replaces traditional vector inner product operations with neural networks to achieve matrix factorization and calculate user ratings for items, and provide personalized recommendations to users. NCF is a general framework based on neural networks for recommendation, consisting of two parts: generalized matrix factorization (GMF) and multi-layer perceptron (MLP). It has the linear modeling ability of generalized matrix factorization and the nonlinear modeling ability of multi-layer perceptron, which can better learn the latent relationship between users and items. In generalized matrix factorization, users and items are encoded into user and item vectors, and the linear latent relationship between users and items is described through the inner product operation of user and item vectors. In a multi-layer perceptron, users and items are encoded and neural networks are used to learn the deep nonlinear latent relationships between users and items. Generalized matrix factorization has linear modeling ability, while multi-layer perceptrons have nonlinear modeling ability. The combination of these two models has stronger learning ability and can better learn the deep latent relationship between users and items [13-15].

In the NCF framework, matrix factorization models can be easily extended and learned from data, resulting in a variant of matrix factorization. Therefore, this method is widely used in practical applications. There is also an extension method that uses nonlinear functions to set the nonlinearity of matrix factorization and learn the latent relationship between users and items from logarithmic loss data. By using nonlinear functions, not only can the model be represented nonlinearly, but also its expressive power can be improved, thereby obtaining the nonlinear relationship between users and items and improving the performance of the model. Linear modeling is generally difficult to obtain deep latent relationships between users and items, therefore, matrix factorization is not very effective in improving collaborative filtering. In the NCF framework, multi-layer perceptrons have nonlinear modeling capabilities and great flexibility, enabling them to learn the deep latent relationships between user and item. Multi-layer perceptron is different from generalized matrix factorization, which can only use the inner product of fixed user and item features. A multi-layer perceptron can gradually reduce the number of neurons from the input layer to the output layer, using fewer neurons at the higher level to learn more abstract features of users and items. Multi-layer perceptrons generally have multiple hidden layers, each with a different number of neurons, so that they can process datasets and complete various tasks. Multi-layer perceptrons have nonlinear modeling capabilities and can learn nonlinear relationships in data. The combination of nonlinear and linear models can further improve the performance of the model.

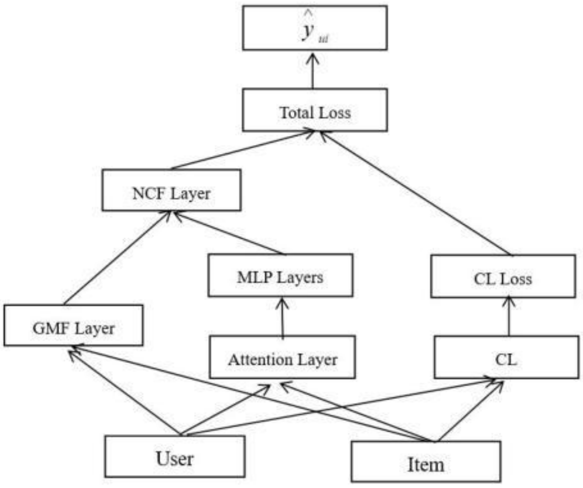


Fig. 3.1: NCF-AC model framework

3. Methodology.

3.1. Problem Definition. Generally, $U = \{u\}$ is represented as a set of users, $I = \{\mathrm{i}\}$ is represented as a set of items, and the historical interaction between user u and item i is represented asy_{ui}.

$$\mathrm{y}_{ui} = \begin{cases} 1 & ,\ \text{interation between u and i} \\ 0 & ,\ \text{otherwise} \end{cases} \tag{3.1}$$

where y_{ui}=1 represents the interaction record between user u and item i, and y_{ui}=0 represents the interaction record that do not exist between user u and project i. The number of user sets U is represented by m, the number of item sets I is represented by n, and the historical interaction matrix between users and items is denoted as Y, where Y is a matrix of $\mathrm{m} \times \mathrm{n}$. Recommendation algorithms mainly predict user ratings for items without interaction based on the historical interaction records between users and items, and recommend high rated items without interaction to users. This process can be abstracted into the form of a learning function, which is used to calculate the score of user u for item i. The function is represented as:

$$\hat{y}_{ui} = f(u,i) \tag{3.2}$$

where y_{ui} represents the prediction score of user u for item i, and f represents the prediction function.

3.2. Framework diagram. The neural collaborative filtering recommendation algorithm is a recommendation algorithm that applies neural network technology to collaborative filtering and has achieved great success. However, the neural collaborative filtering recommendation algorithm treats all items equally and does not consider the different impacts of different items on the recommendation results. The neural collaborative filtering recommendation algorithm also did not consider the sparsity of user item interaction data and the long tail effect. Therefore, we integrates attention mechanism and contrastive learning technology based on the neural collaborative filtering recommendation algorithm to improve the recommendation performance of the recommendation algorithm. By using attention mechanisms to assign different weights to different items, more attention is given to important items. Through contrastive learning, transform the original samples to construct new ones, achieve data augmentation, and optimize model performance. The framework diagram of the neural collaborative filtering recommendation algorithm based on attention mechanism and contrastive learning (AC-NCF) proposed in this paper is shown in Figure 3.1. The following will provide a detailed introduction to attention mechanism and contrastive learning.

3.3. Attention mechanism. Attention mechanism is a technique that simulates human attention, which assumes that humans have different levels of attention towards different items, play more attention to important items, and do not play too much attention into unimportant items. The attention mechanism believes that different items will have different impacts on the results, and all items should not be treated equally. Different items need to be given different weights. Attention mechanism can help us select important and useful information from a large amount of information, play more weight into these information, and improve our work efficiency. In recommendation algorithms, adopting attention mechanisms can better handle the interaction between users and items, and focus more attention on important users and items, and improve the accuracy of recommendation results, and achieve personalized recommendations. The attention mechanism can learn from user behavior which items are more important to the user and which items are not important to the user, in order to better recommend items that the user is interested in [16-17]. The calculation of attention in this paper is shown in formula 3.3, which assigns different weights to different vectors to better match the user's preference for items. This can improve the accuracy of recommending items to users and better meet their personalized needs.

$$\hat{a}_{u,i} = V^T \mathrm{ReLU}(W_a[p_u; q_i] + b_a \ ) \tag{3.3}$$

where W_a and W_b are the weight matrices and bias vectors from the input layer to the hidden layer, respectively. v_r is the attention weight vectors from the hidden layer to the output layer. $[p_u; q_i]$ representing the concatenation of two feature vectors. ReLU is the activation function.

The attention mechanism can capture the attention of user u to item i and assign different weights to different items. When performing feature fusion, the final interaction features between users and items are learned through p_u and q_u. Adopting attention mechanism, it is possible to better learn the latent relationship between users and items, and improve the recommendation performance of the model.

3.4. Contrastive learning. Contrastive learning belongs to self-supervised learning and has been widely applied in fields such as natural language processing, image processing, and computer vision. Contrastive learning constructs variants of the samples that are similar to the samples through transformation processing, adds new samples to the sample space. The enhanced samples should be as close as possible in space, and different samples should be as far apart as possible in space. The core of contrastive learning is data augmentation. In the field of recommendation systems, user and item features are high-dimensional sparse data, and there is a connection between user and item. The interaction data between users and items is a positive sample, and in general, there are relatively few positive samples. Therefore, constructing more positive sample data through contrastive learning is a key issue in contrastive learning [18-19]. The contrastive learning used in this paper is to add random noise to positive samples to construct new samples and achieve data augmentation. For a sample e_i , calculate the sample using formulas 3.4 and 3.5 to construct a new sample.

$$e_i' = e_i + \triangle_i' \tag{3.4}$$

$$e"_i = e_i + \triangle"_i \tag{3.5}$$

where add noise vectors $\triangle_i'$ and $\triangle"_i$ to the original sample, following $\|\triangle\|_2 = \epsilon$ and $\triangle = \bar{\triangle} \odot sign(e_i)$, $\bar{\triangle}$ conforms to a standard normal distribution.

Add noise to the original sample. Rotate the original sample at a small angle, with each rotation corresponding to a new sample. By rotating multiple times, multiple new samples can be obtained, achieving data augmentation. Because the original sample is rotated at a relatively small angle, the resulting sample maintains a certain difference from the original sample after rotation, while also retaining the majority of the information of the original sample. The noise added to the original sample is random and different. The calculation of contrastive loss L_{cl} is shown in formula 3.6.

$$L_{cl} = \sum_i -log \frac{exp(\ e_i^{'T} e_i^{''} /t)}{\sum_j exp(\ e_i^{'T} e_j^{''} /t)} \tag{3.6}$$

where $1/t$ is a constant, T represents the transposition of vectors, i represents a positive sample, j represents a negative sample. The optimization of contrastive loss is to reduce the cosine similarity between different samples obtained through the transformation of the original sample, making the sample distribution more uniform. By optimizing the contrastive loss, similar samples can be made closer in the vector space, while different samples can be further away in the vector space, maximizing useful information from contrastive learning and improving model performance.

3.5. Model optimization. The loss function of the algorithm proposed in this paper consists of two parts, one of which is the loss function L_n of the neural collaborative filtering recommendation algorithm after adding attention mechanism, and the other is the loss function L_{cl} of contrastive learning. The optimization objective of the algorithm proposed in this paper is the sum of loss L_n, loss L_{cl} and L2 regularization.The loss function of the algorithm proposed in this paper consists of two parts, one of which is the loss function L_n of the neural collaborative filtering recommendation algorithm after adding attention mechanism, and the other is the loss function L_{cl} of contrastive learning. The optimization objective of the algorithm proposed in this paper is the sum of loss L_n, loss L_{cl} and L2 regularization.

Loss L_n: This is a paired loss that maximizes the difference in scores between positive and negative samples. Perform matrix factorization on the user rating matrix and optimize it using Bayesian maximum posteriori probability. The loss calculation is shown in formula 3.7.

$$L_n = -\sum_{i\in B} ln(\ y_p - y_n) \tag{3.7}$$

where Y_P represents the score of the positive sample, Y_n represents the score for negative samples, B represents the set of samples.

After obtaining the loss L_n and loss $L - cl$, we combine them for joint learning, as shown in formula 3.8.

$$L = L_n + L_{cl} + \lambda \|\theta\| \tag{3.8}$$

where λ is the regularization coefficient, which is a hyperparameter used to adjust the severity of punishment. The larger the value, the greater the severity of punishment. Adding regularization terms $\lambda \|\theta\|$ is to prevent overfitting.

4. Experimental results and analysis. In this section, we conduct experiments on public datasets to verify the recommendation performance of our proposed algorithm. These experiments mainly answer the following questions:

RQ1: How is the overall performance of our proposed recommendation algorithm compared to other recommendation algorithms?

RQ2: What is the impact of the key parameters of our proposed recommendation algorithm on recommendation performance?

RQ3: What is the impact of adding attention and contrastive learning modules to the neural collaborative filtering recommendation algorithm on recommendation performance?

Firstly, we will introduce the datasets, evaluation metrics and comparison algorithms we use, and then answer the above questions separately. Analyze the overall performance of the algorithm proposed in this paper, analyze the impact of key parameters on recommendation performance, and verify the impact of different modules on recommendation performance through ablation experiments.

4.1. Datasets. In order to verify the recommendation performance of the recommendation algorithm proposed in this paper, experiments were conducted on two public datasets, and the detailed information of these two public datasets is as follows.

1. Movielens dataset: The Movielens dataset is a commonly used machine learning dataset used to evaluate the performance of recommendation algorithms. This dataset is collected through an online movie rating website, which includes user data, movie data, and user rating data for movies. Regardless of the specific rating given by the user to the movie, a score of 1-5 will be uniformly marked as 1 (with interaction), and in other cases, it will be marked as 0. The data is preprocessed. We deleted user data with less than 20 interactions between users and movies.

2. Pinterest dataset: The Pinterest dataset is also a commonly used dataset for evaluating the performance of recommendation algorithms. It is a public and large-scale image recommendation dataset. This dataset contains approximately 50000 users and 1.58 million interaction information between users and items. In order to ensure the quality of the dataset, we also preprocessed the dataset, retaining only user data with 20 or more interactions between users and items.

4.2. Evaluation metrics.

1. HR is a commonly used metric to evaluate the performance of recommendation algorithms, used to evaluate the hit rate of recommendation systems. HR represents the proportion of items that users actually like in the recommendation list. The range of HR values is generally between 0 and 1. The higher the HR value, the more accurate the recommendation.
2. NDCG is the normalized discounted cumulative gain, which is also a commonly used indicator to evaluate the performance of recommendation algorithms. NDCG considers the impact on the ranking order of recommended items by users. Items with higher rankings have higher gains, while items with lower rankings need to compromise their gains. Items with different ranking orders have different contributions to recommendation performance, with the top items having a greater impact and the bottom items having a smaller impact. The range of NDCG values is between 0 and 1, and the higher the value, the better the recommended performance. NDCG is a comprehensive evaluation metric for the recommendation performance of recommendation algorithms.

4.3. Comparison algorithms.

1. Item KNN: This is a recommendation algorithm based collaborative filtering, which obtains a set of similar items to the target item based on the user's historical behavior records of the item. Based on the set of similar items, the algorithm estimates the user's rating on the target item. The algorithm is simple and efficient.
2. BPR: This is a recommendation algorithm based matrix factorization, which learns the latent features of users and items through Bayesian analysis, optimizes using stochastic gradient descent, and then recommends to users based on the ranking results of items. The algorithm has advantages in selecting a very small amount of data for recommendation in massive amounts of data.
3. GMF: This is a generalized matrix factorization recommendation algorithm that represents users and items as a low dimensional user matrix and item matrix through matrix factorization. Then, the user matrix and item matrix are dot products, and item recommendations are made based on the dot product calculation results. The algorithm is a relatively effective method in both explicit and implicit feedback recommendations.
4. MLP: This is a recommendation algorithm based on neural networks, which has a hidden layer or multiple hidden layers. There are multiple neurons in the hidden layer, and the next layer receives input from the previous layer. The layers are converted through nonlinear activation functions. The algorithm has good recommendation performance, but its training process is relatively time-consuming.
5. NCF: This is a recommendation algorithm that applies neural networks to matrix factorization, consisting of two parts: generalized matrix factorization and multi-layer perceptron. It has linear and nonlinear modeling capabilities and can better learn the deep latent relationship between users and items, with good recommendation performance. The algorithm is a neural network-based collaborative filtering algorithm proposed on implicit feedback data, and is a widely used recommendation algorithm.

4.4. Analysis of experimental results. In order to verify the effectiveness of the neural collaborative filtering recommendation algorithm based on attention mechanism and contrastive learning proposed in this paper, experiments were conducted on two public datasets with other benchmark algorithms to compare the recommendation performance. In the experiment, this paper sets the number of layers for the multi-layer perceptron to 3, the vector size to 64, the temperature coefficient to 0.1, the Top-K to 10, and the training frequency to 20. Other benchmark algorithms are set according to the original references. The experimental results are shown in Table 4.1. It can be seen from Table 1 that on the Movielens dataset, the recommendation algorithm proposed in this paper has a higher evaluation metric HR than other benchmark algorithms, and a higher evaluation metric NDCG than other benchmark algorithms. This indicates that adding attention

Table 4.1: Performance Comparison

Dataset	Movielens		Pinterest	
	HR	NDCG	HR	NDCG
ItemKNN	0.701	0.429	0.862	0.536
BPR	0.687	0.418	0.858	0.534
GMF	0.712	0.431	0.860	0.542
MLP	0.720	0.436	0.864	0.547
NCF	0.731	0.442	0.872	0.553
NCF-AC	0.754	0.463	0.915	0.578

mechanism and contrastive learning on the basis of neural collaborative filtering recommendation algorithms is effective in improving the recommendation performance of recommendation algorithms. For Pinterest dataset with a large amount of data, the recommendation algorithm proposed in this paper also performs well in terms of recommendation performance, with higher evaluation metrics HR and NDCG than other benchmark algorithms. This indicates that the recommendation algorithm proposed in this paper is also effective for other datasets, with wide adaptability and good generalization ability. Compared to the Movielens dataset, the Pinterest dataset is a relatively large dataset, and having more samples can enable the algorithm to better learn the representation of users and items, and better filter out items that users may be interested in. The experimental results fully demonstrate that attention mechanism can enable users to pay more attention to important items, better predict user behavior, and recommend more suitable items to users. Contrastive learning can optimize models and improve algorithm performance by changing samples, increasing sample data, and enhancing data representation. Therefore, based on the neural collaborative filtering recommendation algorithm, integrating attention mechanism and contrastive learning can improve the recommendation performance of the algorithm proposed in this paper to a certain extent.

4.5. The impact of parameters on performance. In the recommendation algorithm proposed in this paper, the temperature coefficient τ is an important parameter. If the temperature coefficient is set relatively large, the distribution of negative samples will be smoother, which will cause the model to treat all negative samples equally, thereby affecting the performance of the model. If the temperature coefficient is set relatively small, the model will pay special attention to those hard negative samples, which may cause the model to think that those negative samples may be potential positive samples, which will make it difficult for the model to converge and the model's generalization ability will be poor. The temperature coefficient τ has a significant impact on the performance of the model, therefore, the setting of the temperature coefficient τ plays an important role in the model. Generally, appropriate temperature coefficient values can be determined through experimental analysis. In order to obtain the appropriate temperature coefficient, we set the temperature coefficients separately $\tau = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$, experiments were conducted, and the experimental results are shown in Figure 4.1 and Figure 4.2. From the experimental results, it can be seen that in the Movielens dataset, as the temperature coefficient increases, the performance of the model gradually decreases. When the temperature system is equal to 0.1, the performance of the model is optimal. When the temperature system is equal to 0.1, the performance of the model is optimal. In the Pinterest dataset, the experimental results are basically similar to those in the Movielens dataset. As the temperature coefficient continues to increase, the performance of the model gradually decreases. When the temperature coefficient is equal to 0.2, the performance of the model is optimal. When the temperature coefficient is equal to 0.2, the performance of the model is optimal. The size and sparsity of different datasets vary, and the temperature coefficient is not entirely the same. Through experimental analysis, it is generally found that the effect is better when the temperature coefficient is set in the range of 0.1 to 0.2.

4.6. Ablation experiment. In order to verify the performance impact of adding attention module and contrastive learning module on the recommendation algorithm based on neural collaborative filtering, we conducted ablation experiments on the Movielens dataset and Pinterest dataset. We analyzed the impact of

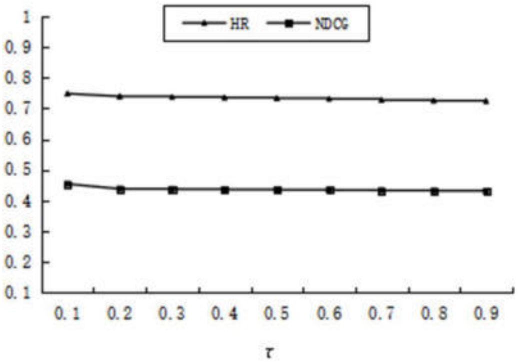


Fig. 4.1: The impact of temperature coefficient τ on performance in Movielens dataset

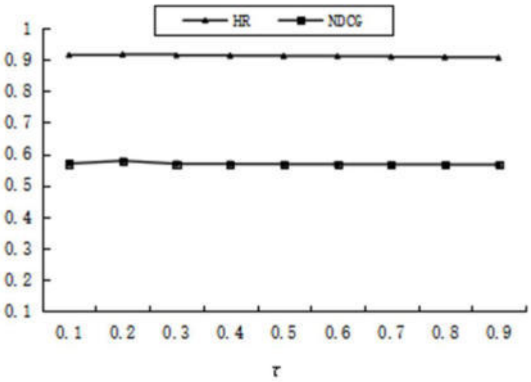


Fig. 4.2: The impact of temperature coefficient τ on performance in Pinterest dataset

different modules on recommendation performance through ablation experiments, and the experimental results are shown in Figures 4.3 and Figures 4.4. We first add only an attention module to the neural collaborative filtering recommendation algorithm, without adding a contrastive learning module, represented as NCF-A. From the experimental results, it can be seen that on the Movielens dataset and Pinterest dataset, both the evaluation metrics HR and NDCG have improved recommendation performance, validating that adding the attention module can improve the performance of recommendation algorithms.Then, we conducted the following experiment by adding only a contrastive learning module and not an attention module to the neural collaborative filtering recommendation algorithm, represented as NCF-C. From the experimental results, it can be seen that on the Movielens dataset and Pinterest dataset, both the evaluation metrics HR and NDCG have improved recommendation performance, validating that adding the contrastive module can improve the performance of recommendation algorithms. From Figure 4.3 and Figure 4.4, it can also be seen that by adding both attention and contrastive learning modules to the neural collaborative filtering recommendation algorithm, the recommendation performance is optimal in terms of evaluation metrics HR and NDCG. Through ablation experiments, we can conclude that using attention mechanism to adjust the weights of different items, focusing more attention on important items, and using contrastive learning to add new samples for data augmentation can effectively improve the recommendation performance of recommendation algorithms.

5. Conclusion. In this paper, we propose a neural collaborative filtering recommendation algorithm based on attention mechanism and contrastive learning, which integrates attention mechanism and contrastive learning on the basis of the neural collaborative filtering recommendation algorithm. Different items have varying degrees

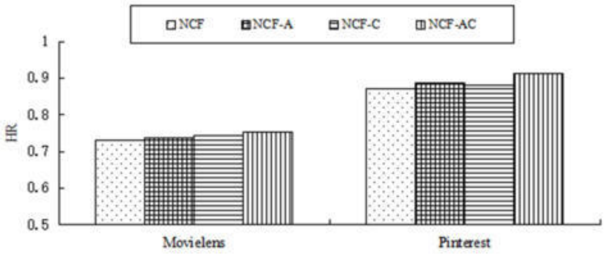


Fig. 4.3: HR values in the ablation experiments

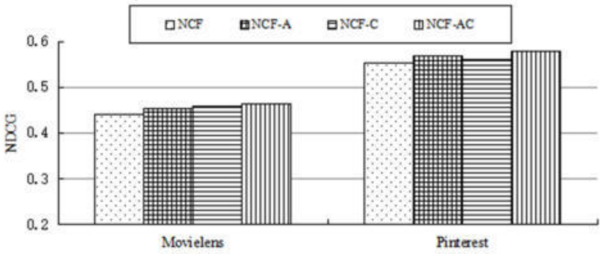


Fig. 4.4: NDCG values in the ablation experiments

of impact on recommendation results, and attention mechanisms adjust the weights of different items based on their varying degrees of impact on recommendation results. In response to the problem of data sparsity, contrastive learning performs transformation operations on samples, increases the number of samples, enhances data representation, and improves model performance by reducing the distance between similar samples in the projection space through contrastive learning loss. Extensive experiments have been conducted on public datasets, and the experimental results show that the algorithm proposed in this paper is effective. Compared with other benchmark algorithms, the recommendation performance has been improved to a certain extent.

REFERENCES

[1] He, X., Pan, J., Jin, O. & Others Practical lessons from predicting clicks on ads at Facebook. *Proceedings Of The Eighth International Workshop On Data Mining For Online Advertising.* pp. 1-9 (2014)

[2] Wang, S., Li, X. & Sun, F. Research review on personalized news recommendation technology. *Computer Science And Exploration.* **14**, 18-29 (2020)

[3] Yu, M., He, W. & Zhou, X. Overview of recommendation systems. *Computer Applications.* **3**, 1-16 (2023)

[4] Goldberg, M., Nichols, F. & Terry, D. Using collaborative filtering to weave an information tapestry. *Communications Of The ACM.* **35**, 61-70 (2002)

[5] Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer.* **42**, 30-37 (2009)

[6] He, X., Liao, L., Zhang, H. & Others Neural collaborative filtering. *International World Wide Web Conferences Steering Committee.* pp. 173-182 (2017)

[7] Chen, L., Wu, L. & Hong, R. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. *Proceedings Of The AAAI Conference On Artificial Intelligence.* **34**, 27-34 (2020)

[8] Kang, L., Chang, D., Jian, H. & Others Self-supervised group graph collaborative filtering for group recommendation. *Association For Computing Machinery.* pp. 69-77 (2023)

[9] Zheng, Y., Gao, C., Chang, J. & Others Disentangling long and short-term interests for recommendation. *Association For Computing Machinery.* pp. 2256-2267 (2022)

[10] Liu, M., Wang, W. & Yang-Xi, L. AttentionRank+: a graph-based recommendation combining attention relationship and multi-behaviors. *Chinese Journal Of Computers.* **40**, 634-648 (2017)

[11] Wang, Y. & Shang, G. A deep collaborative filtering recommendation algorithm that integrates attention mechanism. *Computer Engineering And Applications.* **55**, 8-14 (2019)

[12] Wang, S., Cao, L. & Liang, H. Attention-based transnational context embedding for next-item recommendation. *The Thirty-*

Second AAAI Conference On Artificial Intelligence (AAAI-18). **32**, 2532-2539 (2018)

[13] Qiu, J., Tang, J., Ma, H. & Others Deepinf: Social influence prediction with deep learning. *Proceedings Of The 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 2110-2119 (2018)

[14] He, X., Pan, J., Jin, O. & Others Practical lessons from predicting clicks on ads at Facebook. *Proceedings Of The Eighth International Workshop On Data Mining For Online Advertising*. pp. 1-9 (2014)

[15] Covington, P., Adams, J. & Sargin, E. Deep neural networks for YouTube recommendations. *Proceedings Of The 10th ACM Conference On Recommender Systems*. pp. 191-198 (2016)

[16] You, Y., Chen, T., Sui, Y. & Others Graph contrastive learning with augmentations. *Advances In Neural Information Processing Systems*. **33** pp. 5812-5823 (2020)

[17] Xia, X., Yin, H., Yu, J. & Others Self-supervised hypergraph convolutional networks for session-based recommendation. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **35**, 4503-4511 (2021)

[18] Tang, W., Ren, Z. & Han, F. Collaborative convolutional dynamic recommendation network based on attention mechanism. *Journal Of Automation*. **47**, 2438-2448 (2021)

[19] Wu, Q., Zhang, H., Gao, X. & Others Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. *Proceedings Of The World Wide Web Conference*. pp. 2091-2102 (2019)

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: Apr 29, 2024
Accepted: Jun 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 201–211, DOI 10.12694/scpe.v26i1.3476

TIME SERIES DATA ANALYSIS AND MODELING OF MACHINE LEARNING METHODS IN LIMB FUNCTION ASSESSMENT

XUEYING DUAN[*]

Abstract. In response to the current needs of patients with limb dysfunction, with the goal of safety, real-time, non-invasive, and intelligent rehabilitation assessment, and with limb dysfunction patients as the research object, the author uses intelligent perception technology to obtain rehabilitation data of patients, fully utilizing the advantages of the data itself, and is committed to achieving rehabilitation training and muscle fatigue assessment for limb dysfunction patients. The author developed an assessment model for limb function evaluation using the Dynamic Time Warping K-Nearest Neighbor (DTW-KNN) algorithm and a Long Short-Term Memory (LSTM) neural network-based evaluation model. Based on the experimental findings, it was demonstrated that DTW-KNN effectively categorizes and assesses the rehabilitation motions of upper limbs during elbow flexion under various completion scenarios. Patients have the flexibility to conduct independent and effective upper limb rehabilitation training at home using the upper limb functional rehabilitation system, without any constraints of time or space. By enabling physicians to promptly modify the rehabilitation plan, the system significantly addresses the limitations of conventional upper limb rehabilitation approaches, lowers the medical expenses associated with stroke upper limb rehabilitation, and helps mitigate the shortage of rehabilitation specialists. Utilizing the developed upper limb functional rehabilitation system, the author gathered a set of inertial sensing data on upper limb rehabilitation movements, showcasing prominent temporal features. Consequently, to address this issue, the author proposes the use of Long Short-Term Memory (LSTM) neural network - a recurrent neural network (RNN) with superior temporal data processing capabilities. Based on the multi-dimensional inertial sensing data collected by the upper limb rehabilitation system, the author constructs a recurrent neural network classification model. The model can accurately classify and evaluate different types of upper limb rehabilitation movements under varying completion scenarios. The experimental results indicate that: The overall classification accuracy of DTW-KNN for elbow flexion, elbow flexion&forearm abduction, and shoulder flexion in upper limb rehabilitation movements is 71.8%, 47.9%, and 68.8%, respectively. It was observed that the classification accuracies of LSTM neural network model were 98.2%, 93.3%, and 95.1%, respectively. This marks a notable improvement in the classification accuracy of LSTM neural network model compared to DTW-KNN, with an increase of 26.4%, 45.4%, and 26.3%, respectively. LSTM has a significant advantage over DTW-KNN in terms of classification time, with less classification time.

Key words: Limb rehabilitation, Machine learning, Sensing data, Timing, neural network

1. Introduction. Limb dysfunction refers to clinical pathological changes in the limbs that are not controlled by thinking. Effective rehabilitation training and muscle fatigue monitoring are the basic treatment plans for patients with limb dysfunction, aimed at avoiding disuse atrophy of the patient's limb muscles and improving the body's immunity. At present, medical resources are limited, rehabilitation costs are high, and there are a large number of patients with limb dysfunction. Traditional rehabilitation training models have low efficiency and lack a comprehensive evaluation system for limb rehabilitation training [1]. With the emergence of intelligent rehabilitation, the key to tracking and managing patient rehabilitation is to efficiently and accurately identify and obtain patient rehabilitation training actions, and to conduct real-time muscle fatigue assessment of patient rehabilitation to ensure the safety of rehabilitation training.

At present, there is a huge gap between the medical and health service system and the health needs of residents, and the supply-demand contradiction of medical and health services is becoming increasingly prominent, seriously affecting the healthy development of society [2]. With the continuous intensification of population aging, the total demand for medical services will still maintain a high level. However, due to the imperfect structure of the medical system and the lack of high-quality human resources, the service supply capacity is seriously lagging behind. The growth rate of health technicians and practicing physicians during the same period is clearly unable to meet the annual number of diagnosis and treatment and hospital admissions.

Muscle fatigue is a phenomenon in which the body's muscles are subjected to work activities, resulting in the depletion of energy and substances in the body, affecting muscle energy supply and leading to a decrease

[*] Department of Information Engineering, Jilin Police College, Jilin, China, 130117 (dxyls123456@163.com)

in muscle output power. Muscle fatigue in patients with limb dysfunction is influenced by many factors, and some fatigue cannot be detected through subjective feelings [3]. If patients neglect their muscle fatigue status for a long time and engage in high-load rehabilitation training, it is likely to cause muscle damage. Therefore, real-time evaluation and monitoring of the patient's muscle fatigue level is necessary. Therefore, based on the current limited rehabilitation resources, high rehabilitation costs, and a large number of people in need of rehabilitation, the author will design a limb rehabilitation system based on deep learning and multi-mode sensing data for patient rehabilitation training and muscle fatigue assessment, in order to improve the efficiency of patient rehabilitation training and ensure the safety of patient rehabilitation training.

2. Literature Review. At present, perception of rehabilitation movements, assessment of limb motor function, and assessment of limb fatigue are the most important and widespread needs in intelligent rehabilitation. Within this landscape, entity relationship extraction methods built upon deep learning primarily encode language units of various scales using low-dimensional word vectors, and then uses neural network models such as convolution and loop to achieve automatic learning and extraction of relevant features. Therefore, the author will use technologies such as the Internet of Things, sensors, and artificial intelligence to study rehabilitation motion perception, limb movement function assessment, and limb fatigue assessment. With the comprehensive effects of these three aspects, it will help optimize existing rehabilitation resources, improve patient rehabilitation outcomes, effectively alleviate the shortage of rehabilitation physicians, and is not limited by time and space, with broad application prospects [4]. Xiuli, L. I. et al. observed the effects of upper limb motor games on cognitive function, upper limb motor function, and daily living activities in stroke patients with mild cognitive impairment. Upper limb motor games can promote the recovery of cognitive function, upper limb motor function, and daily living activities in stroke patients with mild cognitive impairment [5].

Intelligent rehabilitation, as a branch of smart healthcare, can achieve a close integration of engineering and medicine, and has the characteristics of strong knowledge professionalism, complexity, and diversity. During the rehabilitation process, the patient's level of limb motor function will constantly change. Real time assessment of limb function can provide effective information for professional physicians and provide reference basis for the optimization of patient rehabilitation training plans in the future. Swarnakar, R. et al. investigated the potential of artificial intelligence and machine learning in assessing, diagnosing, and creating customized treatment plans for individuals with movement disorders. They utilized wearable sensors, virtual reality, augmented reality, and robotic devices to facilitate accurate motion analysis and implement adaptive neural rehabilitation techniques. Additionally, remote rehabilitation powered by artificial intelligence allows for remote monitoring and consultation. Nonetheless, it is imperative for healthcare professionals to interpret the information derived from artificial intelligence and prioritize patient safety. Despite being in the early stages, the effectiveness of artificial intelligence and ML in rehabilitation medicine will be determined through continued research [6].

In recent years, while machine learning has attracted widespread attention from various sectors of society, it has also made significant breakthroughs in the field of rehabilitation. Compared to traditional criteria for evaluating limb motor function, machine learning based methods for evaluating limb motor function are more real-time and accurate. Tang Jinyu et al.'s study observed the clinical efficacy of rehabilitation care in patients with lower limb fractures and its value in preventing complications. In the experiment, patients with lower limb fractures received rehabilitation care and achieved significant clinical efficacy. The fracture healing time of the patients was significantly shortened, the lower limb motor function and knee joint function were significantly improved, the psychological resilience of the patients was increased, and the incidence of complications was significantly reduced [7]. The study by Yaxian, Z. et al. investigated the effects of different intensities of wearable lower limb rehabilitation robot training on walking function, lower limb motor function, balance function, and functional independence in stroke patients. Wearable lower limb rehabilitation robot training may help improve walking function, lower limb motor function, balance function, and functional independence in stroke patients, and high-intensity training may be more effective [8].

Despite the above research, certain issues persist, including: (1) The limb rehabilitation action recognition method solely employs one type of sensor and fails to merge multi-sensor data for rehabilitating action perception, leading to amplified error noise and reduced precision; (2) Due to the high cost of equipment, reliance on a single data source for measurement, and the inability to ensure accurate motion evaluation, the research falls short in meeting the long-term, high-quality rehabilitation needs of patients; (3) The demanding specifications

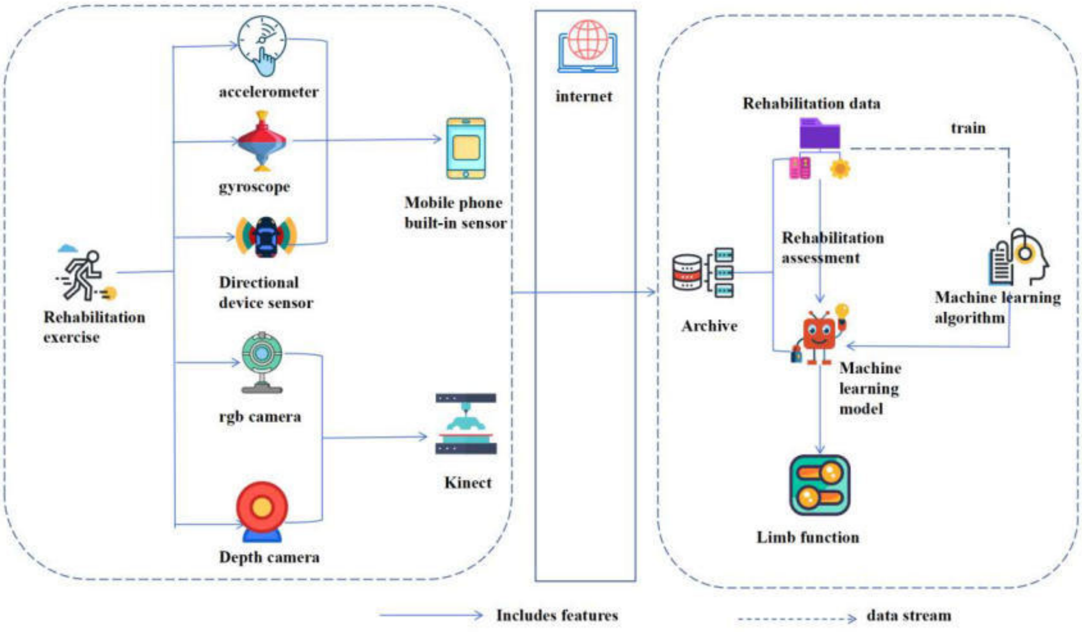


Fig. 3.1: Schematic diagram of limb function evaluation model

for rehabilitation training equipment make it impractical for home use and limit its potential to be accessible to a vast number of patients with limb dysfunction. Therefore, in the context of intelligent rehabilitation, the author collected upper limb rehabilitation movements of patients through low-cost and easily accessible multimodal sensors, and studied the fusion of multimodal sensor data and the perception and evaluation mechanism of rehabilitation movements using machine learning algorithms, achieving low-cost and high-precision rehabilitation action evaluation.

3. Limb function assessment based on machine learning methods . In response to the limitations of existing rehabilitation methods, the author takes multi-modal perception data from rehabilitation training as the research object and explores a limb function evaluation method based on deep learning algorithms. In order to improve the performance of this study, the dynamic time distortion-k nearest neighbor (DTW-KNN) algorithm was selected as a reference to evaluate the LSTM algorithm. Furthermore, the accuracy and efficiency of modeling results between single and multimode data are compared[9]. In addition, the author also conducted a comprehensive discussion on the modeling results.

3.1. Construction of limb function assessment model. A limb function evaluation model that integrates mobile Internet, artificial intelligence, and multi-mode sensors is developed by the author. The structure of this model is divided into two key components: a data collection segment where users directly participate, and a server-based data analysis module. The schematic diagram of the limb function assessment model is shown in Figure 3.1.

Limb movement data is primarily obtained through the built-in inertial sensors found in both mobile devices and Kinect devices. The inertial sensors integrated into mobile devices encompass acceleration sensors, gyroscopes, and directional sensors. On the other hand, Kinect devices utilize RGB cameras and depth cameras for motion sensing purposes. To conduct upper limb movement training, it is necessary for the patient to hold a smartphone equipped with an inertial sensor and select an appropriate standing position in front of the Kinect device before commencing the training session. Upon initiating the training, the smartphone experiences spatial displacement and variations in angles as the patient performs movements with their upper limbs. Real-time data on different upper limb training movements performed by the patient can be monitored and collected

Table 3.1: Display of Original Part Data

a_x	a_y	a_z	ω_x	ω_y	ω_z	Heading angle	Pitch angle	Roll angle	angle
-0.486	0.751	-0.468	-0.04	0.03	0.03	340.6	-48.9	-46.3	1.52
-0.464	0.757	-0.446	-0.04	-0.01	0.02	339.9	-48.7	-46.2	2.44
-0.455	0.763	-0.457	-0.03	0.05	-0.05	340.4	-48.8	-45.9	4.50
-0.464	0.755	-0.467	-0.02	-0.01	0.03	339.5	-48.5	-45.8	4.98
-0.476	0.746	-0.468	-0.03	0.05	-0.05	340.8	-48.9	-46.0	5.97
-0.436	0.770	-0.471	-0.02	0.11	-0.05	339.9	-48.9	-46.2	6.98

through the smartphone's integrated acceleration sensor, gyroscope, and directional sensor [10]. In the inertial sensor mode, the author successfully collected the patient's movement data. Additionally, during upper limb motor training, Kinect somatosensory devices are utilized to capture data on the movement of the patient's limbs. The built-in RGB camera and depth camera of Kinect devices respectively obtain two-dimensional image data and image depth data of patients. Two types of data were preliminarily fused in Kinect, and patient limb joint somatosensory pattern data was obtained. The sensor-collected data from both modes is sent to the server over the Internet for further processing. Following this, the server integrates a variety of machine learning algorithms with diverse pattern data to create several machine learning models for assessing upper limb motions. [11].Once the models are built, inputting the patient's movement data allows for obtaining evaluation results on their upper limb movements. Employing various machine learning models can produce a variety of different outcomes.

3.2. Data preprocessing. Based on the built-in inertial sensors and Kinect of smartphones, the author gathers data on upper limb rehabilitation movements. Subsequently, the collected data on limb rehabilitation is classified and evaluated using DTW-KNN and LSTM algorithms. The inbuilt inertial sensor of the smartphone can determine the sequence data of the patient's upper limb movements by referring to the three-axis coordinate system ofthe phone, including the acceleration values a_x, a_y, and a_z measured by the accelerometer, the angular velocity values ω_x, ω_y, and ω_z measured by the gyroscope when the phone rotates around the three-axis, and the roll angle α, pitch angle β, and heading angle γ measured by the direction sensor. There are a total of 9 types of sequence data. The built-in RGB camera and depth camera of Kinect devices obtain patient limb joint somatosensory mode data. Taking the author's experimental movement of elbow joint flexion as an example, the obtained elbow joint angle . The three joint points of the shoulder, elbow, and hand are all located in the spatial plane, and their range of motion is on the negative half axis of the z-axis in the vector graph. Therefore, the angle between the space vectors ES and EH can be directly used to measure the angle of the elbow joint, and the calculation process is shown in formulas (3.1-3.3):

$$ES = (S_x - E_x, S_y - E_y, S_z - E_z) \tag{3.1}$$

$$EH = (S_x - H_x, S_y - H_y, S_z - H_z) \tag{3.2}$$

$$cos\theta = \frac{ESEH}{|ES||EH|} \tag{3.3}$$

Therefore, the author collected a total of 10 multimodal sequence sensing data, and the raw data collected is shown in Table 3.1.

In this study, there was inevitably a temporal difference in each upper limb movement performed by patients with upper limb dysfunction, so the length of the raw data collected for each upper limb movement was also different. Firstly, the missing data is filled in using anormal distribution, and action sequencesthat are shorter than the specified lengthare extended to ensure that the mean andvariance of the sequence data remain unchanged after interpolation [12]. Secondly, randomly delete action sequences with data entries exceeding

Table 3.2: Action Information

Brunnstrom staging	Action labels	Number of action groups
Phase IV	bad	27
Phase V	in	27
Phase VI	good	27

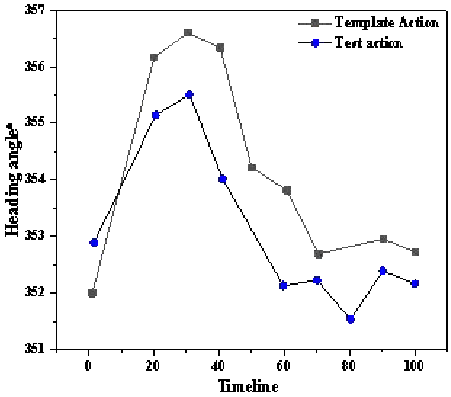


Fig. 3.2: Template Action and Test Action Sequence Trajectory Matching

the specified length, ensuring that the action sequences are of equal length and placed in the LSTM algorithm model. In addition, the collected multimodal sensing data has significant differences in numerical range, and normalization processing is needed to ensure that the data is between 0 and 1, ensuring that it is within the same level. The data normalization is shown in formula 3.4:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.4}$$

where X is the raw data, X_{max} and X_{min} are the maximum and minimum values of the data, respectively.

Due to the different rehabilitation stages and functional states of the upper limbs, the quality of rehabilitation actions performed by patients varies. As shown in Table 3.2, the author evaluated the rehabilitation actions of Brunnstrom IV patients as bad, Brunnstrom V patients as medium, and Brunnstrom VI patients as good. 27 sets of actions were collected for each type of data, totaling 81 sets of actions.

3.3. Construction of a limb function evaluation model based on DTW-KNN. The DTW algorithm, also known as dynamic time warping, is a method used to measure the similarity between two time series of different lengths. It uses the idea of dynamic programming to calculate the similarity between two sequences by stretching and compressing time series.

As shown in Figure 3.2, capture the sampling points from 0 to 100 for the template action and test action, and zoom in on some data segments within the red dashed box. The DTW algorithm performs local point-to-point matching on the trajectories of two time series data to minimize the sum of cumulative distances between the two sequences, thereby comparing the similarity between two non equal length sequences. The black curve N represents the template action sequence, the red curve M represents the test action sequence, and $N_1 - N_{15}$ and $N_1 - N_{16}$ represent the data points on the two sequences, respectively. Calculate the Euclidean distance between two data points to obtain the distance matrix C. The correspondence between points in two sequences can be expressed as formula 3.5:

$$f(k) = (f_n(j), f_m(k)) \tag{3.5}$$

Taking the heading angle in Figure 3.2 as an example, $f_n(k)$ and $f_m(k)$ represent the range of heading angle values from -180 ° to 180 °. The value of k is the number of sensor data collected by the patient for a set of rehabilitation training actions, with a range of values from 1 to S. According to the distance matrix C, the cumulative distance matrix D can be obtained. The solution value for the cumulative distance matrix D is $d_f(N, M)$, and the minimum value is $DTW(N, M)$, as shown in formulas (3.6-3.7).

$$d_f(N, M) = \sum_{k=1}^{S} d(f_n(k), f_m(k)) \tag{3.6}$$

$$DTW(N, M) = mind_f(N, M) \tag{3.7}$$

In this study, the acceleration values a_x, a_y, and a_z measured by the accelerometer, the angular velocity values ω_x, ω_y, and ω_z of the three-axis rotation measured by the gyroscope, the angle data α, pitch angle β, heading angle γ measured by the directional sensor, and the angle data collected by Kinect were a total of 10 dimensional sequence data. Therefore, the sum of the distances from each dimension is the distance between two rehabilitation action sequences.

The K Nearest Neighbor (KNN) algorithm is currently a commonly used data classification method. The author selects a K value of 11 and inputs the cumulative distance between two rehabilitation action sequences obtained by the DTW algorithm into the KNN classifier to complete the classification evaluation of patient rehabilitation training actions [13].

3.4. Construction of a limb function evaluation model based on LSTM . Each LSTM cell has 3 inputs and 2 outputs. The inputs include the multidimensional sensing data input x_t

3.5. Experimental setup and evaluation indicators. The author's experiment used word vectors trained by the Word2Vec algorithm for word embedding, with a dimension of 300. Based on prior knowledge, the K parameter was set to 20, the alpha parameter was set to $4e^{-2}$, and the optimal values of other parameters were determined using a grid search algorithm on the dataset. Finally, the optimal results were achieved in the 55th to 60th iteration rounds. The optimal parameter settings for the model are shown in Table 4.2. at the current moment, the LSTM cell output and the cell state h_{t-1} at C_{t-1} the previous moment, and the outputs include the output value h_t and cell state C_t. Currently, the unit status represents the transmission process of information. The LSTM network mainly relies on the unit state C and the current output h for model training. W_f,W_i, W_c, and W_o represent weights, b represents bias terms, σ is the sigmoid function, and tanh is the hyperbolic tangent function, as shown in formulas (3.8-3.9).

$$Sigmoid(z) = \frac{1}{1 + e^{-z}} \tag{3.8}$$

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.9}$$

LSTM cells contain three basic structures, namely output gate, input gate, and forget gate. The three gating units of LSTM each have independent weight matrices and skewing parameters, which can change the connection weight and skewing for each time step data. This design is conducive to avoiding gradient vanishing and explosion, and is suitable for processing long-term multi-dimensional sensing data [14]. The function of the forget gate is to forget unnecessary information and control the magnitude of the forgotten input x_t and the previous hidden layer output h_{t-1}, as shown in formula (3.10):

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3.10}$$

The function of the input gate is to save new information to the cell state. Firstly, the input gate determines the information in the cells that need to be updated by using an S-shaped function. As shown in formula (3.11):

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3.11}$$

Table 4.1: Detailed indicators of elbow joint flexion A/B/C completion test movements

index	Class A test action	B-class test action	C-class test action
Accuracy	86.3%	100%	53.2%
recall	78.2%	57.2%	92.3%
Specificity	95.8%	100%	65.2%
F1-Score	82.7%	72.4%	67.6%

After determining the information that needs to be updated, use a tanh function to generate a vector and obtain alternative update content as shown in formula (3.12):

$$C'_t = tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3.12}$$

The final author combines these two parts to obtain a new cell state, as shown in formula (3.13):

$$C_t = f_t * C_{t-1} + i_t * C'_t \tag{3.13}$$

The responsibility of the output gate is to determine the final output content based on the cell state. Firstly, LSTM uses a sigmoid function to determine which part of the cell state to output:

$$o_t = \sigma(W_o[h_t, x_t] + b_o) \tag{3.14}$$

After determining the output part of the cell state, the cell state will be processed by tanh and multiplied by the result of the sigmoid function to obtain the final output result, as shown in formula (3.15):

$$h_t = o_t * tanh(C_t) \tag{3.15}$$

In this study, the input information $x_t = (x_{t1}, x_{t2}, \cdots, x_m)$ represents an n-dimensional feature vector generated by normalizing the original data, which is the multi-dimensional elbow joint flexion motion data obtained by multi-mode sensors. Set the time step to S, take the top S data of the current data to be classified, and obtain $S * n$ dimensional data. The author integrates these data into an input matrix and sends it into the recurrent neural network model. Each time, the data sample $x_t(t = 1, 2, \cdots, s)$ at time t is placed, and a total of S times are placed in the loop. h_{t-1} is the output result of the model at time $t - 1$, which is fed into the hidden layer at time t to obtain the classification prediction result h_t of the action sequence at time t. The LSTM algorithm excels at extracting features from time series and integrating multi-dimensional sensor action data. The simplest method is to treat multi-dimensional sensor action data as a complete multi-dimensional time series. In the author's multivariate time series modeling using multi-dimensional elbow joint flexion motion data as input, updating the state of each time node in the multi-dimensional sensing data is crucial, and a neural network model suitable for multiple input variables is needed. The LSTM algorithm can perfectly solve this difficult problem.

4. Experimental results. In order to evaluate the effectiveness of different pattern data and machine learning algorithms in rehabilitation action classification, it is necessary to compare the accuracy of classification results obtained using different algorithms for different pattern data. This article used DTW-KNN and LSTM algorithms to model the collected data and analyzed the classification results obtained.

(1) The classification of DTW-KNN. It is as follows.

For elbow flexion movements, the overall classification accuracy of DTW-KNN is 71.8%.

The overall classification accuracy of DTW-KNN for elbow flexion and forearm abduction movements is 47.9%.

For shoulder flexion movements, the overall classification accuracy of DTW-KNN is 68.8%. Tables 4.1 to 4.3 provide detailed indicators of the completion of these movements.

Table 4.2: Detailed indicators of elbow flexion and forearm abduction A/B/C completion test movements

index	Class A test action	B-class test action	C-class test action
Accuracy	34.5%	56.8%	0%
recall	56.7%	74.3%	0%
Specificity	63.2%	56%	100%
F1-Score	42.2%	65.2%	0%

Table 4.3: Detailed indicators of shoulder joint flexion A/B/C completion test movements

index	Class A test action	B-class test action	C-class test action
Accuracy	46.85%	92.7%	100%
recall	89.1%	63.4%	63.2%
Specificity	66.5%	90.2%	100%
F1-Score	62.2%	71.6%	77.2%

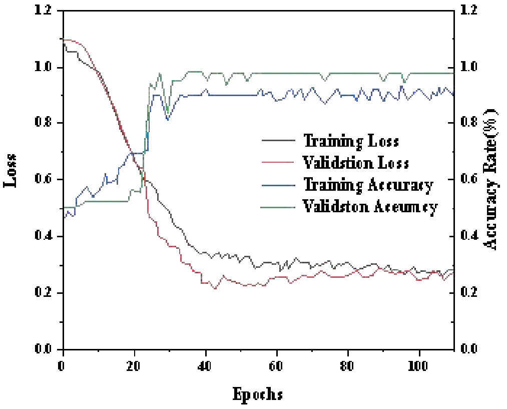


Fig. 4.1: Classification error and accuracy of elbow flexion LSTM neural network model

Table 4.4: Classification Efficiency Analysis of Two Classification Algorithms for the Completion of Three Upper Limb Rehabilitation Actions

Algorithm& Efficiency	action	elbow flexion	Elbow flexion& forearm abduction	Shoulder joint flexion
DTW-KNN Accuracy	71.8%	47.9%	68.8%	
time consuming	127.98s	93.28s	95.18s	
LSTM Accuracy	98.2%	93.3%	95.1%	
time consuming	3.18s	5.41s	3.25s	

(2) LSTM model error and accuracy. Figures 4.1 to 4.3 show the classification errors and accuracy of LSTM neural networks corresponding to elbow flexion, elbow flexion&forearm abduction, and shoulder flexion.

Based on the classification results of three types of upper limb rehabilitation exercises completed by different classification models, as well as the comprehensive efficiency analysis in Table 4.4, from the above analysis, it can be seen that the DTW-KNN algorithm has good overall classification results for the three types of upper

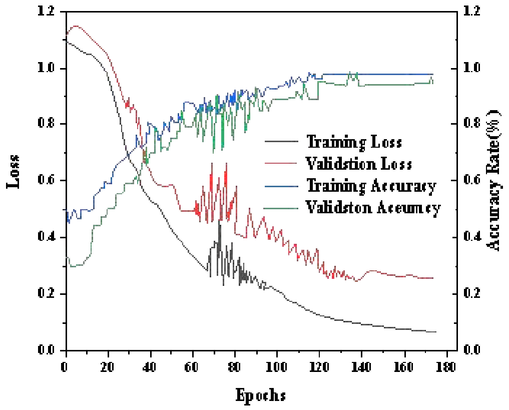


Fig. 4.2: Classification error and accuracy of the LSTM neural network model for elbow flexion and forearm abduction

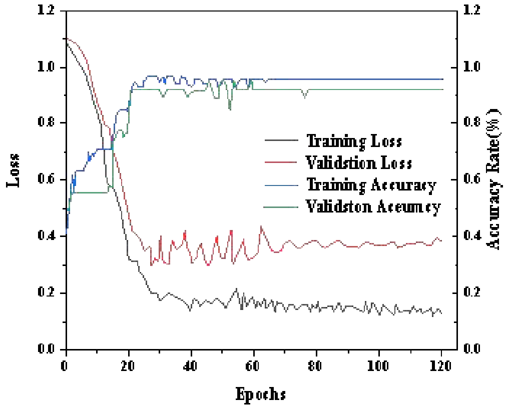


Fig. 4.3: Classification error and accuracy of LSTM neural network model for shoulder flexion

limb rehabilitation movements, but there are differences in accuracy between different rehabilitation movements. The classification accuracy of elbow flexion and shoulder flexion rehabilitation exercises is excellent, but the classification accuracy of elbow flexion and forearm abduction movements is relatively low, only 47.9%. From Table 4.1, it can be seen that the accuracy, recall, and F1 Score of the elbow flexion and forearm abduction movements with completion status C are all 0, indicating that the DTW-KNN classification model cannot accurately classify the movement with completion status C. In addition, among the three types of upper limb rehabilitation exercises, the DTW-KNN classification model provides a more accurate classification of actions with a completion state of B. In elbow flexion, the classification results for actions with a completion state of C are relatively average, while in shoulder flexion, the classification results for actions with a completion state of A are not as good as the corresponding classification results for actions with a completion state of B and C.

The overall fitting of the LSTM neural network model is good. The classification accuracy of LSTM for the above three types of upper limb rehabilitation movements is 98.2%, 93.3%, and 95.1%, respectively. Compared to DTW-KNN, the classification accuracy of neural network models has improved by 26.4%, 45.4%, and 26.3%, respectively. As the number of model iterations increases, the error stabilizes at a relatively low level. In addition, LSTM takes much less time for classification than DTW-KNN, making it more advantageous.

The author takes the common rehabilitation training action of elbow joint flexion as an example, and the results show that the rehabilitation training platform based on multi-mode sensor technology can ensure that patients complete rehabilitation training at home and accurately obtain patient rehabilitation training motion data [15]. In limb function assessment tasks, the evaluation effect of multi-mode sensing data is better than that of single-mode sensing data. The DTW-KNN algorithm performs well in low dimensional data, while in high-dimensional data, the LSTM algorithm performs better in accuracy and time overhead compared to the DTW-KNN algorithm.

5. Conclusion. In response to the current needs of patients with limb dysfunction, with the goal of safety, real-time, non-invasive, and intelligent rehabilitation assessment, and with limb dysfunction patients as the research object, the author uses intelligent perception technology to obtain patient rehabilitation data, fully utilizing the advantages of the data itself, and is committed to achieving rehabilitation training and muscle fatigue assessment for limb dysfunction patients. An efficient and accurate method for evaluating the completion of limb retraining movements is crucial for patients' home rehabilitation. Considering the significant temporal nature of the inertial sensing data collected by the author for limb rehabilitation movements, and the good performance of recurrent neural networks in solving sequence data problems, the author considers using LSTM, which has good performance in processing sequence data, in order to solve practical classification problems. This article provides a detailed explanation of the modeling process for constructing a classification model for action completion based on the inertial sensing data collected from limb rehabilitation movements in this study and the LSTM neural network. In addition, under the guidance of a rehabilitation therapist, the author collected three limb rehabilitation movements: elbow flexion, elbow flexion&forearm abduction, and shoulder flexion under different completion conditions, and conducted DTW-KNN and LSTM comparative experiments, the overall fitting of the LSTM neural network model is good. The classification accuracy of LSTM for the above three types of upper limb rehabilitation movements is 98.2%, 93.3%, and 95.1%, respectively. The classification accuracy of LSTM neural network model has improved by 26.4%, 45.4%, and 26.3% compared to DTW-KNN, respectively. LSTM has a significant advantage over DTW-KNN in terms of classification time, with less classification time.

Real time and accurate assessment of limb function can enable rehabilitation physicians to timely understand the patient's health status, optimize the patient's rehabilitation training plan, and effectively improve the quality of patient rehabilitation. At present, rehabilitation treatment resources are limited, rehabilitation costs are high, and the number of people in need of rehabilitation is huge. In addition, traditional rehabilitation training evaluation methods rely on the professional knowledge of physicians, with strong subjectivity and low accuracy. The completion of the patient's rehabilitation training plan after discharge is not satisfactory, and there is a lack of a comprehensive rehabilitation training evaluation system. In response to this situation, the author has developed a rehabilitation training platform based on multi-mode sensor technology, where patients can receive high-quality rehabilitation training at home according to the rehabilitation plan formulated by physicians. The rehabilitation training platform obtains multi-mode sensing data of patients' rehabilitation through Internet technology, inputs them into the limb function evaluation model, and effectively feeds back the evaluation results of rehabilitation training to patients and doctors in real time, so as to improve the rehabilitation quality and training enthusiasm of patients.

As the effectiveness of patient rehabilitation training continues to improve, the patient's rehabilitation training plan needs to be synchronously optimized. The existing human-machine collaboration methods have poor effectiveness, generally based on single factor considerations, and the efficiency of utilizing expert knowledge is low. They cannot effectively combine rehabilitation medical data with expert knowledge to apply to patient rehabilitation decision support. With the emergence of massive rehabilitation information, clinical diagnosis and treatment doctors and rehabilitation physicians are facing a huge challenge. Clinical diagnosis, treatment, and rehabilitation plans for patients can only be formulated based on subjective personal experience, making it difficult to obtain and reuse past rehabilitation treatment plans, resulting in the waste of medical resources. Therefore, a medical decision support system can be established by combining the massive medical data generated by the rehabilitation industry with the limb movement function results evaluated by patients. Medical decision support systems use cutting-edge technologies such as machine learning and artificial intelligence to conduct in-depth analysis and inference of diverse medical structures and related professional knowledge, thereby assisting doctors in optimizing rehabilitation plans and predicting rehabilitation risks, improving treatment efficiency and service quality for patients.

Acknowledgement. Topic: Jilin Provincial Department of Science and Technology, Jilin Provincial Science and Technology Development Plan Project - Research on Key Industrial Core Technologies - New Generation Information Technology Field - New Generation Network Communication Technology, Big Data and Artificial Intelligence

Project name: Research and development of limb function assessment system based on AI and machine vision

REFERENCES

[1] Quiroz, D., Greene, J. M., Limb, B. J., & Quinn, J. C. . (2023). Global life cycle and techno-economic assessment of algal-based biofuels. Environmental Science & Technology: ES&T(31), 57.

[2] Chen Liya,Zhong Juanxu ,Liu Zhongqi,Chao hope, Juanxu, Z., Zhongqi, L., & Pan, C. . (2023). Analysis of the value of graded nursing under caprini risk assessment in the prevention of lower limb deep vein thrombosis after total hysterectomy. Chinese and Foreign Medical Research, 21(20), 118-121.

[3] Lu, Y. H., Fu, Y., Shu, J., Yan, L. Y., & Shen, H. J. . (2023). Application of cross-migration theory in limb rehabilitation of stroke patients with hemiplegia. The World Journal of Clinical Cases, 11(19), 4531-4543.

[4] Zhang, Y., Wang, D., Wang, D., Yan, K., Yi, L., & Lin, S., et al. (2023). Motor network reorganization in stroke patients with dyskinesias during a shoulder-touching task: a fnirs study. Journal of Innovative Optical Health Sciences, 16(06).

[5] Xiuli, L. I., Shan, L. I., Mengchen, F., & Fubiao, H. . (2023). Effects of upper limb exergames on functional recovery in stroke patients with mild cognitive impairment. Chinese Journal of Rehabilitation Theory and Practice, 29(1), 98-103.

[6] Swarnakar, R., & Yadav, S. L. . (2023). Artificial intelligence and machine learning in motor recovery: a rehabilitation medicine perspective. The World Journal of Clinical Cases, 11(29), 7258-7260.

[7] Tang Jinyu& Shuhong, L.(2023).Analysis of the application effect of rehabilitation nursing in patients with lower limb fractures. Chinese and Foreign Medical Research, 21(18), 80-84.

[8] Yaxian, Z., Zhiqing, T., Xinting, S., Rongrong, W., Tianhao, L., & Hao, Z. . (2023). Effects of different intensity of wearable lower limb rehabilitation robot-assisted training on lower limb function after stroke. Chinese Journal of Rehabilitation Theory and Practice, 29(5), 497-503.

[9] Umunnah, J., Adegoke, B., Uchenwoke, C., Igwesi-Chidobe, C., & Alom, G. . (2023). Impact of community-based rehabilitation on quality of life and self-esteem of persons with physical disabilities and their family member. Global Journal of Health, 7(2), 87-93.

[10] Park, D., Son, K. J., & Kim, H. S. . (2023). Chronic phase survival rate in stroke patients with severe functional limitations according to the frequency of rehabilitation treatment. Archives of physical medicine and rehabilitation, 104(2), 251-259.

[11] Zhu, M., Guan, X., Li, Z., He, L., Wang, Z., & Cai, K. . (2023). Semg-based lower limb motion prediction using cnn-lstm with improved pca optimization algorithm.Journal of Bionic Engineering, 20(2), 612-627.

[12] Wang, S. . (2023). A parkinson's disease diagnosis approach for nonequilibrium gait data. International Journal of Modeling, Simulation, and Scientific Computing, 14(05).

[13] Feng-Hua, Y. U., Ju-Chi, B., Zhong-Yu, J., Zhong-Hui, G., Jia-Xin, Y., & Chun-Ling, C. . (2023). Combining the critical nitrogen concentration and machine learning algorithms to estimate nitrogen deficiency in rice from uav hyperspectral data. Journal of Integrative Agriculture, 22(4), 1216-1229.

[14] Yang, D., Wang, L., Yuan, P., An, Q., Su, B., & Yu, M., et al. (2023). Cocrystal virtual screening based on the xgboost machine learning model. China Chemical Express, 34(8), 398-403.

[15] Zhang, Q., Zheng, W., Song, Z., Zhang, Q., Yang, L., & Wu, J., et al. (2023). Machine learning enables prediction of pyrrolysyl-trna synthetase substrate specificity. ACS Synthetic Biology, 12(8), 2403-2417.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Mar 11, 2024
Accepted: Apr 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 212–219, DOI 10.12694/scpe.v26i1.3536

DEVELOPMENT AND APPLICATION OF SPORADIC MATERIAL INVENTORY OPTIMIZATION MANAGEMENT SYSTEM BASED ON ARTIFICIAL INTELLIGENCE

QIAN LI[*], WENJUN HOU[†], YATONG WU[‡], YAQI HOU[§], AND JIAN ZHU[¶]

Abstract. This paper introduces the overall framework of the intelligent electrical sporadic materials management system based on the ubiquitous Internet of Things. The system includes storage equipment and an intelligent management terminal. It makes optimal design for storage equipment and intelligent management terminals. The dynamic modeling of a three-layer supply chain system of power materials is constructed and composed of manufacturers, suppliers, and raw materials. The inventory control strategy of each link in the supply chain of power materials under different cooperation levels is studied by introducing the corresponding adjustment variables. Through the experiment and analysis of the system, it is proved that the system has good performance in label management, resource management, inventory management, disposal management, system management and essential management. The number of concurrent users, response speed, stability and other aspects of the system are excellent. The database is complete, independent, and secure. And the data is objectively reasonable and repairable. The system can lay a specific technical foundation for intelligent management of electrical sporadic materials.

Key words: Artificial intelligence; Power material optimization; Inventory optimization; Power material supply chain

1. Introduction.

The auxiliary products produced by State Grid Power Company have a lot of uncertainties (such as great demand and large randomness). However, the traditional power supply chain information transmission system is subject to the lagging development of science and technology and the lack of effective control measures, resulting in high internal information concentration and imperfect information transmission mechanisms between different levels. It was found that due to insufficient information sharing and a slow interaction rate among enterprises at each link in the supply chain, information transmission between all links in the network will be significantly affected. Due to the lack of adequate information disclosure, the overall decision-making efficiency and effect of the whole supply chain are greatly restricted. Using new scientific and technological means is essential to effectively manage the existing supply chain and inventory of power materials. Literature [1] builds a supply chain management system based on blockchain, applies this system to the automotive industry and achieves an excellent win-win situation. Literature [2] uses the blockchain method to build a vehicle supply chain traceability system, which effectively alleviates the lack of credit among enterprises in all aspects of the vehicle supply chain. The system improves the security and traceability of vehicle information. Literature [3] introduced blockchain technology in the manufacturing industry and established the concept of ”sharing culture” among enterprises, thus effectively improving the overall service quality of the manufacturing industry. Literature [4] argues that blockchain technology can generate three kinds of value: sharing, security, and smart contracts. These three features are embodied in the supply chain as encryption, consensus mechanisms, and smart contracts. Literature [5] introduced blockchain technology in the production process of dairy products to build distributed transaction records. The system can accurately track important information throughout the supply chain. This project intends to study the intelligent inventory management system of electrical sporadic materials in the ubiquitous Internet of Things environment, aiming to achieve the purpose of interconnection, ubiquitous visualization and intelligent management of electrical sporadic materials.

[*] State Grid Shanxi Electric Power Company Materials Branch, Taiyuan, Shanxi, China, 030021 (Corresponding author, 15333662419@163.com)

[†] State Grid Shanxi Electric Power Company Materials Branch, Taiyuan, Shanxi, China, 030021

[‡] State Grid Shanxi Electric Power Company Materials Branch, Taiyuan, Shanxi, China, 030021

[§] State Grid Shanxi Electric Power Company Materials Branch, Taiyuan, Shanxi, China, 030021

[¶] State Grid Shanxi Electric Power Company Materials Branch, Taiyuan, Shanxi, China, 030021

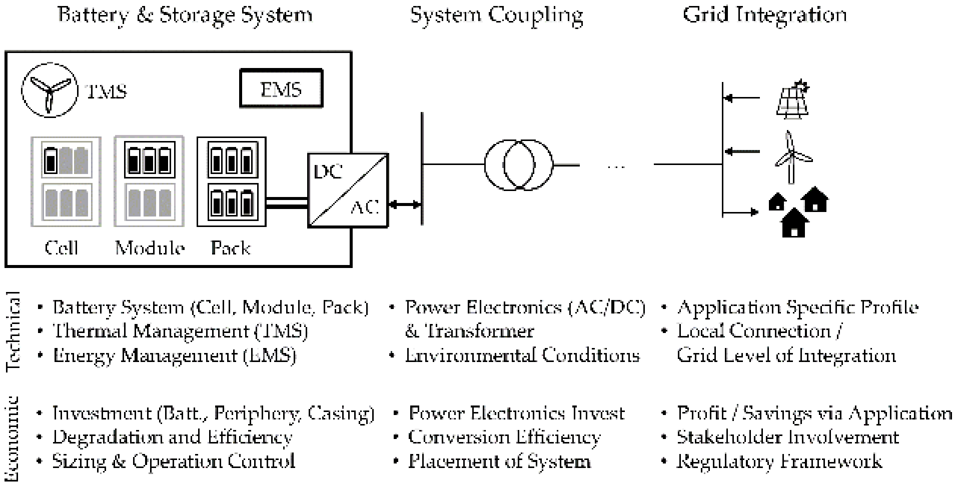

Battery & Storage System
System Coupling
Grid Integration
TMS
EMS
DC
AC
Cell
Module
Pack
Technical
• Battery System (Cell, Module, Pack)
• Thermal Management (TMS)
• Energy Management (EMS)
• Power Electronics (AC/DC) & Transformer
• Environmental Conditions
• Application Specific Profile
• Local Connection / Grid Level of Integration
Economic
• Investment (Batt., Periphery, Casing)
• Degradation and Efficiency
• Sizing & Operation Control
• Power Electronics Invest
• Conversion Efficiency
• Placement of System
• Profit / Savings via Application
• Stakeholder Involvement
• Regulatory Framework

Fig. 2.1: Physical architecture of the electrical sporadic materials inventory management system.

2. Design of optimization management system for electric sporadic materials inventory.

2.1. System Architecture. The physical architecture of the intelligent management system for electric sporadic materials is shown in Figure 2.1 (the picture is quoted in Energies 2017, 10(12), 2107). Storage devices store waste power and waste materials. A load cell is arranged on each pad of the storage device. An intelligent management terminal is in the middle, under the tool [6]. The weighing sensor transmits the weight signal of the electric power device to the intelligent management terminal through the junction box. The intelligent management terminal uses 4 G/5 GNB-not to transmit the weight, position, height and other information to the intelligent management system [7]. Electrical sporadic materials are an intelligent management system that stores and transports waste materials for monitoring.

2.2. Optimal design of transportation and storage devices. Firstly, the whole system is optimized. The aim is to reduce its quality. The original half-fan door is transformed into a one-type door, which can effectively improve power materials' loading and unloading efficiency [8]. The lock structure has been improved to reduce the time required for unlocking and locking. The weighing part on the gate is combined with the whole gate by mounting the weighing element on each base plate. The weight of loading and unloading equipment is significantly reduced, and the total amount of stored electrical energy is increased.

2.3. Optimization design of intelligent management terminal. Information such as the location, quality, elevation and production process of electric power materials are stored intelligently and uploaded to the intelligent management platform of electric power materials [9]. The operation status of storage and transportation equipment is monitored, and parameters are set by using the human-machine interface. It is the central link to the whole process of monitoring and managing power generation equipment. The system comprises weighing, positioning, communication, and power supply modules. The modular architecture allows for easy maintenance and updates. This system separates weighing and other functional modules and the weighing sensor and intelligent management terminal are connected. The weighing sensor is connected to the terminal box, and the terminal box does not need to be opened when the weighing sensor is replaced to facilitate the installation, use, maintenance and update of the weighing sensor [10]. In addition, the spacing of each weighing element also reserves sufficient margin. Each weighing element carries about 10 t, which can ensure the quality of carrying electric energy materials. In the weighing process, the intelligent algorithm directly displays the necessary power materials in the storage and transportation unit on the web page.

Unlike the conventional single-frequency positioning method, the positioning module uses the L1+L5 frequency band. The L5 band has a high signal coding rate. The dual frequency band can effectively suppress the position deviation caused by the atmospheric ionosphere and significantly improve the position accuracy of

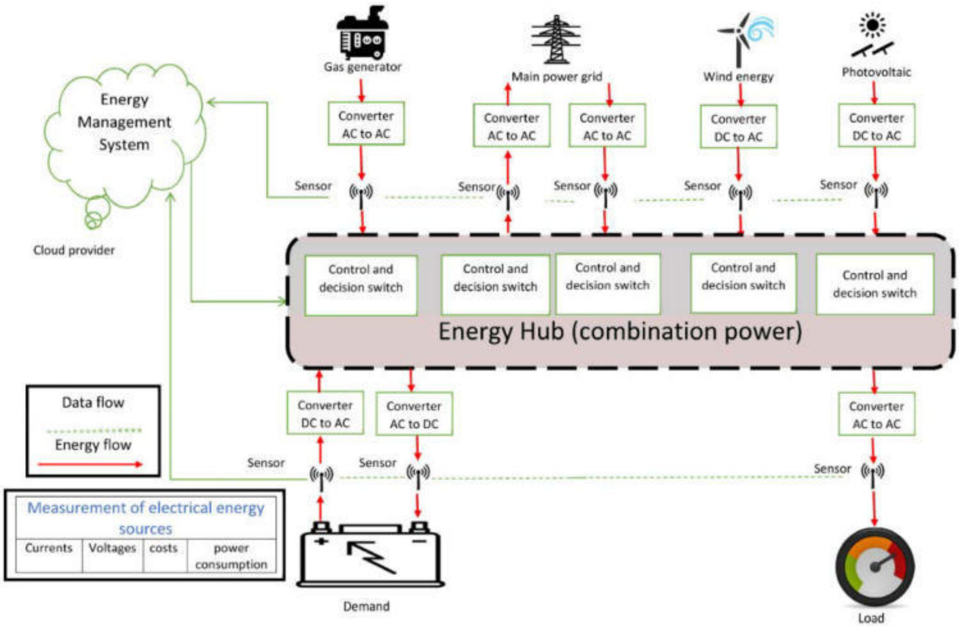


Fig. 2.2: Structure of intelligent power material management system based on ubiquitous Internet of Things.

the satellite. Intelligent management terminals can reach an accuracy of 3-5 meters in open areas. The relative height of the box is calculated by using the difference of air pressure in the height measurement to calculate the floor where the box is located accurately. The intelligent management terminal retains an upgradeable test interface, built-in RS-485, I/O and other ports for easy plugging and maintenance [11]. The 5G Internet of Things adopted in this paper has the characteristics of low power consumption and high anti-interference ability.

The power module uses a safe battery, which can increase the battery's capacity without increasing the housing size [12]. In the process of use, if there is an abnormal phenomenon, the system will send an email to remind the user. LCC batteries have a wide operating temperature range compared to nickel-chromium and lithium batteries.This battery will not spontaneously ignite even if a puncture occurs and charges faster.

2.4. Realization of service functions of the system. The platform can control the state of the whole process, analyze the whole data and make intelligent decisions [13]. The system is divided into three levels: data input layer, platform application layer and background management layer. Figure 2.2 shows the structure of the intelligent management system for electric power materials (the picture is quoted in Appl. Sci.2021, 11(21), 9820).

The core of the power materials management system is to input and manage the electronic label of power materials. The intelligent management process of resources, inventory and processing is realized [14]. Material management is mainly used to manage warehouse resources, such as warehouse number, location, automatic distribution, etc. Inventory management includes inventory count, inquiry, inventory count, warehouse capacity analysis, monthly inventory utilization, storage season and seasonal analysis, etc. Disposal management includes disposal method input, time reminder, plan formulation, approval, and early warning. The parameter setting layer includes the system management and essential management functions. These include name/number, role/position, data/function authorization, and system operation records. Essential management includes the basic functions of warehousing management, such as environmental parameter setting, purchase requirements management, warehousing rules setting, etc. The module also provides supplier/carrier management and transfer document management functions.

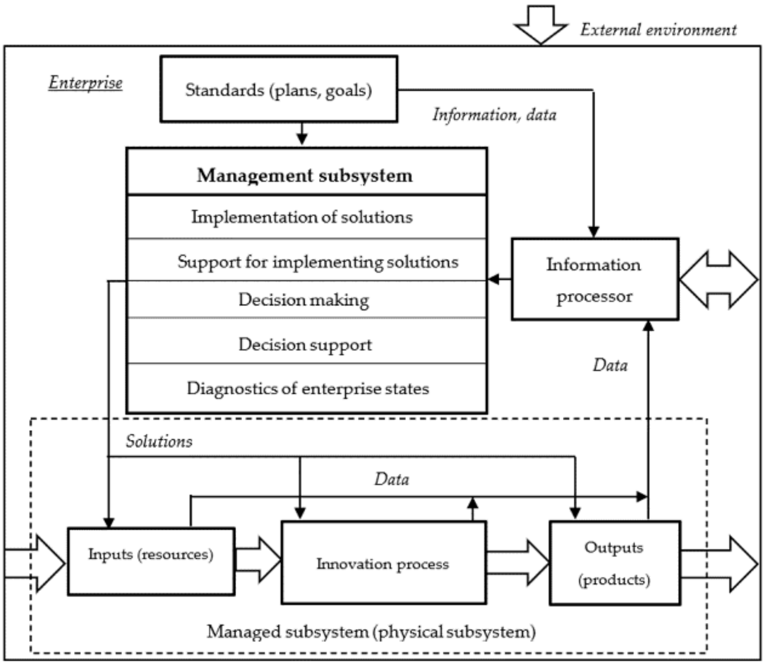


Fig. 2.3: Improved business process for intelligent management of power materials.

2.5. Business process optimization. The steps are the followings:

1. Change the traditional telephone booking mode and use the intelligent power supply equipment management platform to achieve booking.
2. The manually registered electric energy materials in the warehouse are reformed, and the intelligent management terminal is used to input and store electric energy materials.
3. Change the traditional warehouse and construction worker measurement modes and realize the automatic measurement by transportation unit measurement. The data is transmitted to the system through the intelligent management terminal and verified by the system.
4. The warehouse manager's location assignment mode is improved to automatically configure the location according to the optimal algorithm of the location allocation.
5. Change the method of manual inventory of electrical energy materials.

The system provides intelligent decision support for the disposal of power supplies, the optimal allocation of cargo locations, the inventory check of power supplies and other work.

3. Supply chain inventory management based on block database model. If there is no complete collaboration in the block-based database model, and only their inventory costs are shared, they will aim to optimize their inventory and make production and order forecasts. Let the quantity demanded $K(\alpha \leq K \leq \beta)$ be random and equally distributed. So now people have the distribution function $G(K) = \frac{K-\alpha}{\beta-\alpha}$ and the probability density $g(x) = \frac{1}{\beta-\alpha}$, so $\alpha = v - \sqrt{3}\varepsilon, \beta = v + \sqrt{3}\varepsilon$. The expected value is $v = \frac{\alpha+\beta}{2}$ and the standard deviation is $\varepsilon = \sqrt{\frac{(\beta-\alpha)^2}{12}}$. Table 3.1 lists the relevant variables and explanations.

3.1. Optimization of material platform orders. The material platform makes ordering decisions based on demand information shared by the various demand parties in the modular database [15]. The formula for calculating the inventory cost of the material platform can be obtained:

$$W_1\left(C_s\right) = Y_w \int_{Q_s}^{\infty} \left(K - Q_s\right) g_1(K) dx + C_s \int_{0}^{Q_s} \left(Q_s - K\right) g_1(K) dx + Y_s Q_s$$

Table 3.1: *Variables and Definitions.*

Variable	Paraphrase
C_m	Unit price of producer's inventory
C_n	Price of supplier's inventory
C_r	Material platform unit inventory cost
Q_m	Producer's output
Q_n	Quantity ordered by supplier
Q_r	Quantity ordered by material platform
P_m	Manufacturer's manufacturing cost per product
P_n	Supplier's unit price
P_r	Unit price of material platform
P_w	Unit price of material platform
C_z	The unit price of supply chain inventory
E	Total inventory

Taking the derivative of Q_s in formula (3.1) gives:

$$\frac{dW_1\left(C_s\right)}{dQ_s}=-Y_w\left[1-G_1\left(Q_s\right)\right]+C_sG_1\left(Q_s\right)+Y_s$$

When (3.2) =o, the optimal purchase quantity $Q_s^*=\frac{\alpha(C_s+Y_s)+\beta(Y_w-Y_s)}{C_s+Y_w}$ is obtained.

3.2. Supplier's Optimal Ordering Decision. Construct the ordering strategy of the raw materials platform based on the block database. The relationship between inventory costs and suppliers is as follows:

$$W_2\left(C_n\right)=Y_s\int_{Q_n}^{\infty}\left(Q_s-Q_n\right)g_2\left(Q_n\right)dQ_n+C_n\int_0^{Q_n}\left(Q_n-Q_s\right)g_2\left(Q_n\right)dQ_n+Y_nQ_n(3\cdot 3)$$

Taking the derivative of Q_n in (3.3) yields:

$$\frac{dW_2\left(C_n\right)}{dQ_n}=-Y_s\left[1-G_2\left(Q_n\right)\right]+C_nG_2\left(Q_n\right)+Y_n$$

If (3.4) is 0 , the manufacturer's optimal order quantity $Q_n^*=\frac{\alpha(C_n+Y_n)+\beta(Y_s+Y_n)}{C_n+Y_s}$ is obtained. 3.3 Manufacturer's Optimal Output Decision. Manufacturers can schedule production through vendor ordering decisions based on the block database. From the above assumptions, the formula for calculating the manufacturer's inventory cost can be obtained:

$$W_3\left(C_m\right)=Y_n\int_{Q_m}^{\infty}\left(Q_n-Q_m\right)g_3\left(Q_m\right)dQ_m+C_m\int_0^{Q_m}\left(Q_m-Q_n\right)g_3\left(Q_m\right)dQ_m+Y_mQ_m$$

Taking the derivative of Q_m in (3.5) yields:

$$\frac{dW_3\left(C_m\right)}{dQ_m}=-Y_n\left[1-G_3\left(Q_m\right)\right]+C_mG_3\left(Q_m\right)+Y_m$$

If formula (3.6) = o, the optimal product yield $Q_m^*=\frac{\alpha(C_m+Y_m)+\beta(Y_n+Y_m)}{C_m+Y_n}$ of the manufacturer can be obtained.

3.3. Model construction based on independent priority. Block database provides a good platform for information sharing. It enables the enterprises in each link of the supply chain of power materials to quickly obtain reliable information from the supply chain and downstream enterprises [16]. This paper develops a dynamic model based on independent dominance (Figure 3.1 cited in Processes 2021, 9(6), 982).

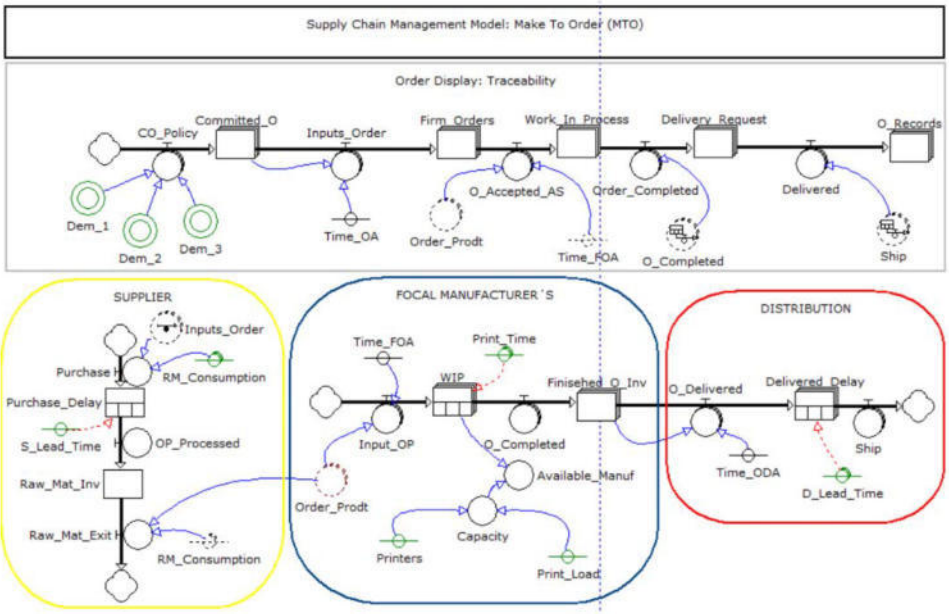


Fig. 3.1: System dynamics model under respective dominant modes.

3.4. Optimal Decision under Smart Contract Association. This project intends to introduce the power material management supply chain model based on blockchain technology to form a close cooperation relationship between producers, suppliers and raw material platforms through the raw material platform to share inventory. In contrast, suppliers share the cost of inventory [17]. Through the block database, the production planning of the supply chain is discussed, so $q_m = Q_m = Q_n = Q_s, C_c = C_s$. When the ordering decisions of these three people are the same, the inventory cost of the entire supply chain can be expressed by the following formula.

$$dW_4\left(C_c\right) = Y_w \int_{q_m}^{\infty} \left(K - q_m\right) g_1(K) dK + C_s \int_{0}^{q_m} \left(q_m - K\right) g_1(K) dK + Y_n q_m$$

Taking the derivative of q_m in formula (3.7) gives:

$$\frac{dW_4\left(C_c\right)}{dq_m} = -Y_w\left[1 - G_1(K)\right] + C_s G_1(K) + Y_n$$

When equation (3.8) = 0, the optimal production product $q_m^* = \frac{\alpha(C_s+Y_n)+\beta(Y_w+Y_n)}{C_s+Y_w}$ can be obtained.

4. Test and analysis. The black box test method is used to test it to check the performance and quality of the ubiquitous IoT power management system developed in this paper. The black box method tests the system software's function by inputting test cases and checking whether the output meets the expectations [18]. The functional test results of the intelligent management system for electric power materials are shown in Table 4.1. The performance test results of the intelligent management system for electric power materials are shown in Table 4.2. The test results of the database security of the intelligent management system for electric power materials are shown in Table 4.3.

5. Conclusion. This paper designs the physical and functional architecture of the intelligent management system of power materials based on the ubiquitous Internet of Things. Finally, an example is given to verify the performance indicators of the intelligent electric power logistics management system under the "ubiquitous Internet of Things" proposed in this paper, which meets users' requirements. The platform has a comprehensive

Table 4.1: System function test results.

Serial number	Test content	Output result	Test result
1	Label management	Consistent with the desired output	Successful test
2	Material management	Consistent with the desired output	Successful test
3	Inventory control	Consistent with the desired output	Successful test
4	Waste disposal	Consistent with the desired output	Successful test
5	System Administration	Consistent with the desired output	Successful test
6	Grass-roots administration	Consistent with the desired output	Successful test

Table 4.2: System performance test results.

Serial number	Test content	Test result
1	Parallel operand	P 1000
2	Data collection rate	95% or higher
3	LAN response time	<100 ms
4	User registration response time	3.6 s or less
5	Main menu response time	3.1 s or less
6	Timeliness of intelligence response	93% or higher
7	Static information integrity	93% or higher
8	Accuracy of information	93% or higher
9	Stability of system	No problems, such as system crash, occurred
10	Database stability	stable

Table 4.3: Results of database security tests.

Serial number	Exam question	Test result
1	Database security	The database is encrypted
2	Database integration degree	Complete database for all business activities to provide data support
3	Database data manageability	It supports the operation of adding, modifying and deleting system data.
4	Database independence	Independence from other systems
5	Database storage and recovery functions	In the event of an error, the database has a complete backup and can be restored quickly

function and a good security guarantee ability, and it can be well adapted to current power material management needs.

6. Acknowledgement. This article is the research result of the 2023 State Grid Shanxi Electric Power Company's technology project ”Research on the Self evaluation Method of Sporadic Materials Based on Artificial Intelligence” (Project code: 52051W230001).

REFERENCES

[1] Oladele, T. O., Ogundokun, R. O., Adegun, A. A., Adeniyi, E. A., & Ajanaku, A. T. (2021). Development of an inventory management system using association rule. Indonesian Journal of Electrical Engineering and Computer Science, 21(3), 1868-1876.

[2] Balkhi, B., Alshahrani, A., & Khan, A. (2022). Just-in-time approach in healthcare inventory management: Does it really work. Saudi Pharmaceutical Journal, 30(12), 1830-1835.

[3] Pasaribu, J. S. (2021). Development of a Web Based Inventory Information System. International Journal of Engineering, Science and Information Technology, 1(2), 24-31.

[4] Bader, D., Innab, N., Atoum, I., & Alathamneh, F. (2023). The influence of the Internet of things on pharmaceutical inventory management. International Journal of Data and Network Science, 7(1), 381-390.

[5] Nozari, H., Ghahremani-Nahr, J., & Szmelter-Jarosz, A. (2023). A multi-stage stochastic inventory management model for transport companies including several different transport modes. International Journal of Management Science and Engineering Management, 18(2), 134-144.

[6] Agboola, F. F., Malgwi, Y. M., Mahmud, M. A., & Oguntoye, J. P. (2022). Development of a Web-Based Platform for Automating an Inventory Management of a Small and Medium Enterprise. FUDMA Journal of Sciences, 6(5), 57-65.

[7] Aris, M. A. M., & Salikon, M. Z. M. (2022). Development of Warehouse Inventory Management System: Pembangunan Sistem Pengurusan Inventori Gudang. Applied Information Technology and Computer Science, 3(1), 529-540.

[8] Thakre, S. (2021). study of the impact of jit (just–in time) in inventory management in the automobile sector India in. IJM, 12(2), 720-730.

[9] Yankah, R., Osei, F., Owusu-Mensah, S., & Agyapong, P. J. (2022). Inventory management and the performance of listed manufacturing firms in Ghana. Open Journal of Business and Management, 10(5), 2650-2667.

[10] Neve, B. V., & Schmidt, C. P. (2022). Point-of-use hospital inventory management with inaccurate usage capture. Health Care Management Science, 25(1), 126-145.

[11] Mahajan, P. S., Raut, R. D., Kumar, P. R., & Singh, V. (2024). Inventory management and TQM practices for better firm performance: a systematic and bibliometric review. The TQM Journal, 36(2), 405-430.

[12] Herlambang, C. A., & Parung, J. (2021). Information system design and inventory management on pharmacy business within ABC-XYZ analysis method. Airlangga Journal of Innovation Management, 2(2), 194-205.

[13] Kihara, B. W., & Ngugi, P. K. (2021). Inventory management systems and performance of public hospitals in Kenya; case of counties under universal Health care programme. International Journal of Social Sciences and Information Technology, 7(2), 66-77.

[14] Preil, D., & Krapp, M. (2022). Artificial intelligence-based inventory management: a Monte Carlo tree search approach. Annals of Operations Research, 308(1), 415-439.

[15] Ali, K., Showkat, N., & Chisti, K. A. (2022). Impact of inventory management on operating profits: Evidence from India. Journal of Economics, Management and Trade, 28(9), 22-26.

[16] Qi, M., Shi, Y., Qi, Y., Ma, C., Yuan, R., Wu, D., & Shen, Z. J. (2023). A practical end-to-end inventory management model with deep learning. Management Science, 69(2), 759-773.

[17] Dey, B. K., & Seok, H. (2024). Intelligent inventory management with autonomation and service strategy. Journal of Intelligent Manufacturing, 35(1), 307-330.

[18] Cahyono, A., & Titisari, M. A. (2023). Implementation of Erp-based Automated Inventory Management System at Sabana Fried Chicken Franchise Company. Journal of Economics and Business UBS, 12(5), 3307-3319.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Mar 18, 2024
Accepted: May 8, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 220–230, DOI 10.12694/scpe.v26i1.3669

TEACHING QUALITY EVALUATION AND IMPROVEMENT BASED ON BIG DATA ANALYSIS

XUEQIU ZHUANG[*]AND MEIJING SONG[†]

Abstract. To address the limitations of Problem-Based Learning (PBL) and to foster student initiative while enhancing teaching quality, the author suggests a novel approach: leveraging big data analysis for teaching quality evaluation and improvement. This method involves conducting diverse and dynamic evaluations, randomly and repeatedly, involving students, teachers, and supervisors. By applying an enhanced Dempster evidence synthesis formula and weights derived from the Analytic Hierarchy Process, the system dynamically calculates each teacher's rating in their respective courses, allowing for continuous improvement. Additionally, personalized feature indicators and teaching quality evaluation metrics are developed to provide a comprehensive assessment. The results indicate that in the coarse evidence set algorithm, is obtained through experience. If is used as the weight alone, the subjectivity is too heavy, and is added for fusion operation, as well as the intervention of experience factor, a balance point between subjectivity and objectivity is found. The final score of 4.2878 was obtained by combining the weights between subjects obtained through Analytic Hierarchy Process, which is consistent with the survey and the public's opinion. This method avoids the deficiency of traditional evidence theory that treats all evidence equally, enhances the ability of information fusion, and obtains more realistic conclusions. Further validated the feasibility and usability of the personalized teaching quality evaluation and improvement model for software engineering.

Key words: Teaching quality, Personalized feature indicators, Apriori, Dempster evidence synthesis, Analytic Hierarchy Process

1. Introduction. Nowadays, the educational concept of ”student-centered” has permeated various fields of educational activities, and improving student learning effectiveness has become a value demand for the development of higher education [1]. The quality of higher education teaching plays a pivotal role in nurturing talented individuals, directly influencing the caliber of graduates produced. Student learning outcomes serve as a tangible indicator of this quality [2,3]. Thus, enhancing teaching quality serves as a crucial avenue for improving talent cultivation in higher education institutions. It aligns with broader educational policies, bolsters the capacity of higher education to contribute to societal progress, drives teaching reforms, facilitates strategic positioning and distinctive operations for universities, elevates the standard of talent cultivation, and ultimately enhances student learning achievements.

As higher education continues to grow and evolve, the demand for effective teaching quality assurance systems that cater to students' expectations and aspirations becomes increasingly imperative. With rising enrollment rates and ever-changing dynamics within the educational landscape, there's a constant influx of new trends and student preferences. This necessitates a responsive approach to meet the evolving needs of students and ensure that their learning outcomes are consistently met. Students have diverse and diverse choices of courses, school curriculum is diverse, teaching equipment is constantly upgraded and updated, and learning resources are abundant. However, no one can accurately know whether students have achieved substantial improvement in their learning effectiveness. Merely evaluating the teaching quality of teachers cannot provide evidence for improving teaching quality. The ultimate measure of teaching quality lies in its impact on student learning outcomes. No matter how comprehensive the teaching content or high the teaching quality, if it fails to address the actual needs of students, it cannot effectively improve their learning effectiveness. Therefore, universities must prioritize student learning outcomes and establish robust teaching quality assurance systems focused on students' growth and development. By ensuring that students truly learn and benefit from their

[*]School of Finance and Economics, Hainan Vocational University of Science and Technology, Haikou, Hainan, 570000, China (Corresponding author, jhh@mail.qtnu.edu.cn)

[†]School of Management, Universiti Sains Malaysia, Minden, Penang 11800, Malaysia

higher education experience, universities can uphold the quality of talent cultivation and pave the way for meaningful teaching reforms [4,5].

2. Literature Review. Data mining can mine or extract valuable patterns or patterns hidden within a large amount of incomplete, noisy, fuzzy, and random data [6,7]. It includes many specific methods, including neural networks, classification, decision trees, regression analysis, clustering, and association rules.

In recent years, more and more scholars have begun to study the application of data mining technology in teaching quality evaluation. Among many methods, clustering, association rules and other algorithms are the most commonly used. Yang, Y. H. et al. conducted a study on the correlation between XAPP (eXtreme Apprenticeship Pedagogical Pattern) and the management of teaching quality, employing the Plan-Do-Check-Act (PDCA) cycle as a framework. Their objective was to enhance teaching methods and quality to achieve superior teaching outcomes. To examine the teaching process of XAPP comprehensively, they established a closed-loop management system focusing on various modules: building the teaching team, designing courses, implementing courses, selecting textbooks, evaluating and adjusting teachers, conducting customer satisfaction surveys, and providing teaching evaluation and feedback. Through an analysis of the PDCA cycle's application in XAPP teaching quality management, they aimed to optimize piano teaching management practices, enhance the XAPP teaching system, refine curriculum design, improve textbook selection, and enhance the theoretical framework [8]. Pertuz, S. et al. proposed a quality evaluation method for use in blended learning and described an application case study of quality evaluation methods for three undergraduate courses in electronic engineering projects [9]. Che, Y. and others investigated the integration of decision trees in physical education teaching within the realm of big data, leading to the development of a comprehensive physical education teaching management system. Recognizing that the quality of physical education teaching significantly influences the advancement of school physical education, they emphasized the critical role of curriculum selection and enrichment in this process. By collecting and refining textbooks, they aimed to enhance teaching resources, stimulate personal growth and reflection, and elevate the artistic and creative aspects of teaching [10]. Zhao, L. and others employed quantitative and data analysis techniques to thoroughly explore the effectiveness and acceptance of the interactive teaching mode in blended English education. Their research aimed to provide a comprehensive understanding of how students perceive and engage with this teaching approach. The quantitative analysis revealed that a significant portion of respondents expressed favorable attitudes towards the interactive teaching mode, indicating a high level of acceptance among students. Only a minority of participants reported dissatisfaction with this teaching model [11].

With the continuous development of network technology, the teaching staff of the Software College is constantly strengthening. Teaching for software engineering must now rely on computer tools. Every PBL mode teaching in the Software College is fully capable of allowing students and teachers to evaluate and monitor the teaching process through computer networks. Based on the network and relying on information technology, we have transformed from questionnaire surveys to online evaluations, utilizing the storage functions of the network and data to achieve dynamic evaluation and timely feedback capabilities in learning. The author constructs a personalized teaching quality evaluation and improvement model for software engineering, which can not only continuously adjust existing evaluation models through data sampling and analysis, but also establish a complete set of personalized learning kits for students, improve the shortcomings of PBL mode, enhance students' initiative, and improve the teaching quality of teachers, enabling them to adapt to students' learning methods.

3. Research Methods.

3.1. Teaching quality indicators.

3.1.1. Theoretical Principles of Teaching Quality Indicators. The author prioritizes addressing the challenge of tailored education by developing personalized teaching quality evaluation criteria, essential for enhancing teaching quality and customizing courses to meet individual student needs. Focused on software engineering, the author aims to establish a comprehensive model for evaluating and improving teaching quality, specifically tailored to this field [12]. The objective of this personalized evaluation approach is to provide guidance for achieving teaching quality objectives in software engineering education. By delineating clear teaching

Table 3.1: Student Evaluation Index System

Top level indicators	Second level indicators
Training Methods	Is there a unique teaching style Is the course case appropriate Is the language clear and the writing neat Whether multimedia devices such as PPTs are effectively used Can theory be combined with practice Is the teaching simple and easy to understand Homework assignment and feedback are reasonable
Teaching content	The teacher's expertise in the field of the course Is lesson preparation sufficient Is the amount of teaching information appropriate Strong logical content, prominent concepts, key and difficult points
Teaching attitude	Attend classes on time without unauthorized rescheduling Passionate and serious teaching Is it better to organize classroom teaching Answering student questions with a serious attitude Harmonious teacher-student relationship and sufficient communication Fair and strict treatment of students
Teaching effectiveness	Gained from this course Attractive teaching by teachers Are you willing to choose other courses from this teacher Teaching while emphasizing the cultivation of student abilities

philosophies and evaluation standards, universities can ensure relevance and practicality in assessing personalized teaching. Ultimately, this approach empowers teachers to concentrate on enhancing teaching quality and fostering students' enthusiasm for learning. Although this is a personalized teaching quality assessment for software engineering, we also hope to adapt to various course types, teaching fields, and teaching methods of various majors in a comprehensive university. So it is required that the personalized teaching quality evaluation system should have both commonality and purposefulness, in order to meet more teaching modes and personalized needs. The evaluation indicators not only need to reflect the qualified standards of teaching quality, but also need to have the ability to improve. The previous teaching evaluation was only based on a simple weight distribution, which could not well meet and reflect the teaching situation of teachers. Not only would it lose the targeted improvement of teaching work, but the classification would also be too detailed, making the evaluation meaningless.

In summary, in order to establish a teaching quality indicator system, it is necessary to combine student evaluation, peer evaluation, and supervision evaluation as the basis. In the teaching process, teachers play a leading role, while students are the main body. The teaching methods, methods, and content are highly integrated in classroom teaching. In the previous text, we have addressed the different factors of cultural background, knowledge level, values, etc. in student evaluation [13-14]. So, by considering the differences in the areas of focus among peers and supervisors, and combining the differences of the above three subjects, we can objectively and fairly reflect the content of teaching quality evaluation and the direction of improvement. There are only three aspects of evaluation with different weights. The evaluation of the process is led by the teacher, but if the evaluation is conducted once every semester or half a semester, it is not significant enough, because the quality of teaching is closely related to the teacher's moral character, cultural level, teaching level, and preparation status, as well as their understanding of the teaching content, teachers can fully prepare before evaluation to cope with it. Therefore, evaluation must have continuity and randomness. At least, student evaluations should be maintained at least 4 times a semester, and peer and supervisory evaluations should be conducted at least 2 times, and conducted randomly.

3.1.2. Establishment of Teaching Quality Index System.

1) The establishment of a student evaluation index system. The establishment of a teaching quality indicator system should widely draw on the evaluation standards of domestic and foreign universities, combine with the actual work of the School of Software at Nanchang University, and use a hierarchical thinking and top-down approach to determine. Based on the previous evaluation standards for teaching quality at Nanchang University, the top-level evaluation indicators should include four aspects: teaching methods, teaching content, teaching attitude, and teaching effectiveness. The second level evaluation indicators are further classified, and the rating content of each indicator is divided into five levels: 5, 4, 3, 2, 1. The specific content of the table is shown in Table 3.1.

Table 3.2: Peer Teacher Indicator System

Peer teacher evaluation indicators
Adequate lesson preparation and complete lesson plans
Suitable material selection and accurate concept
The content is novel and practical
Highlighting key points and clear regulations
Clear language and neat writing
Thinking driven, ability cultivation
Patience in tutoring and timely feedback on homework
Teaching according to individual needs and strong extracurricular interaction
Fair assessment and appropriate rating

Table 3.3: Indicator System for Supervisors

Supervisor evaluation indicators
Basic concepts correct
The basic theory is correct
Highlights of this lesson
Thorough analysis of difficulties
Explaining is inspiring
Clear regulations with strong logical coherence
Clear language and neat writing
Integrating theory with practice and emphasizing the cultivation of abilities
Pay attention to student emotions

The basis for setting the content involved is:

(1) Teaching method: The correct method is of great significance in improving the quality of teaching. So improving teaching methods places the most emphasis on seven second level indicators.

(2) Teaching content: It is the core of classroom teaching, directly related to what students can learn, which has a significant impact on the quality of teaching.

(3) Teaching attitude: Teaching attitude is also an important factor affecting the quality of teaching, and whether a teacher is serious and responsible is related to whether students can effectively absorb and acquire knowledge.

(4) Teaching effectiveness: The effectiveness is the standard for testing whether students have learned knowledge, and the quality of teaching ultimately depends on the feedback of the teaching effectiveness.

2) Establishment of peer teacher evaluation indicators. Because peer teachers are very familiar with classroom teaching content and have the same teaching experience, they have a considerable say in whether the teaching content in the classroom is correct, innovative, and handled properly [15]. However, due to the relatively small number of peer teachers compared to students, the evaluation indicators of peer teachers are shown in Table 3.2, and the evaluation criteria are still 5 levels.

3) Establishment of evaluation indicators for supervisors. Supervisors are the main responsible persons representing the school's supervision of the teaching quality of teachers. They play the role of monitors in the teaching quality system, and their evaluation index system is shown in Table 3.3, which also has 5 different levels of indicators.

All data can be obtained through two aspects:

1. Collect data one by one through a questionnaire survey and enter it into the database.
2. In the computer operation course, you can directly operate on the computer and transfer it to the database through the network.

3.2. Algorithm for evaluating teaching quality.

3.2.1. Improved Dempster evidence synthesis formula. In the evaluation indicators of teaching quality, the indicators involved have non-linear relationships. If we rely solely on experience or treat all indicators equally. The Dempster synthesis rule provides a synthesis formula for these indicators, but in the evaluation of teaching quality, each indicator has different degrees of influence on the final value.

Rough set is an effective tool for handling various incomplete information such as imprecision, inconsistency, and incompleteness.

The specific steps for applying the comprehensive teaching quality evaluation algorithm are as follows:

1. Obtain data on teaching quality evaluation indicators (such as student evaluation indicator system) and form an information system S=(U,A,V,f); Among them, U represents the set of non empty valid objects, which we call the domain; A is the set of all indicators; V is the value range of indicator a, which is V=(5,4,3,2,1) in the teaching quality evaluation system; F is an information function that specifies the attribute values of each object in U.

2. Attribute importance, set $X \subseteq A$ as a subset of attributes and $a \subseteq A$ as an attribute. The importance of a to X is denoted as Sig X(a), and its calculation formula is Equation 3.1:

$$Sigx(a) = 1 - |X \cup \{a\}|/|X| \tag{3.1}$$

The meaning of this definition is to randomly select two objects in U, with a total of 2U selection methods, among them, there are |X| that are indistinguishable under the attribute subset X, and there are $|x| \cup \{a\}|$ cases that are indistinguishable after adding attribute a in X, which are obviously less than or equal to |X|. Therefore, $|X| - |X \cup \{a\}|$ represents the indistinguishable reduction in X due to the addition of attribute a, of course, it is the distinguishable increase, which refers to the number of selection methods that were previously indistinguishable under X but are now distinguishable under $|X \cup \{a\}|$.

Calculate the value of evaluation indicator $C = \{a_1, a_2, \cdots, a_n\}$ using Sig X(a), and each $SigX(a_i)$ is the corresponding attribute importance.

3. Normalize the importance $SigX(a_i)$ of each obtained attribute according to $\lambda_i = SigX(a_i)/\sum_{i=1}^{n} sigx(a_i)$, and this λ_i is the weight value of each second level teaching quality evaluation indicator [16].

4. Determine the confidence level $M_i(A_i), i = 1, 2, \cdots, n$ for each attribute a_i in student, peer teacher, and supervisor $C = \{a_1, a_2, \cdots, a_n\}$ based on college experience. Among them, $\sum_{A \subseteq \theta} m(A) = 1$.

5. Based on the experience of the college, select the appropriate experience factor and calculate the comprehensive reliability $M_i'(A_i)$ of a_i, as shown in equation 3.2

$$M_i'(A_i) = M_i(A_i) \times \theta + \lambda_i \times \sum_{A \subseteq \theta} m(A) \times (1 - \theta) \tag{3.2}$$

where θ=[0,1], a smaller value indicates a greater emphasis on objective weight, and a larger value indicates a greater emphasis on experience.

6. By combining the evidence theory formula, equations 3.3 and 3.4 are obtained

$$m(A) = K^{-1} \times \sum_{\cap A_i \leqslant A} \prod_{1 \leqslant i \leqslant n} M_i'(A_i) \tag{3.3}$$

$$K = 1 - \sum_{n A_i \leqslant A} \prod_{1 \leqslant i \leqslant n} M_i'(A_i) \tag{3.4}$$

Obtain the final score for the evaluation of teaching quality.

3.2.2. Application of Analytic Hierarchy Process in Teaching Quality Evaluation System . After obtaining comprehensive evaluation values from three aspects, we still cannot determine the specific score of the teaching quality level of the course teacher, because the evaluation scores of the three main subjects: Students, peer teachers, and supervisors are not comparable to each other. Therefore, determining the weight

Table 3.4: Comparison of the Importance of Three Main Evaluation Indicators

i j	Student subject	Teacher peer subject	Supervisor subject
Student subject	1		
Teacher peer subject	count backwards	1	
Supervisor subject	count backwards	count backwards	1

Table 3.5: Comparison of the Importance of Three Main Evaluation Indicators

Scale	Definition (comparing subjects i and j)
1	Subject i is equally important as j
3	Subject i is slightly more important than j
5	Subject i is clearly more important than j
7	Subject i is more important than j
9	Subject i is extremely important than j
2,4,6,8	Middle value between two adjacent values
count backwards	The inverse comparison between subject i and j $b_{\bar{j}} = 1/b_{ij}, b_{ii} = 1$

of each subject in the overall evaluation value is a very important issue that must be solved. In actual teaching evaluation, the weight is generally determined based on experience or leadership decisions, but the problem is that the actual weight difference is too large. The final result is inaccurate evaluation. So the objectivity and scientificity of weights directly affect the teaching quality evaluation of the course teachers. Analytic Hierarchy Process (AHP) is an analytical method for multi-objective decision-making, which combines qualitative and quantitative analysis, decomposes elements related to decision-making problems into levels such as objectives, criteria, and plans, and digitizes decision-making thinking. Analytic Hierarchy Process (AHP) constructs a judgment matrix B by comparing the relative importance of factors pairwise, and calculates the weights of the importance orders that are related to each other. This is called hierarchical single ranking. Determine the consistency of the matrix by calculating the eigenvalues and eigenvectors of B. The specific steps are as follows

1. In the teaching quality evaluation system, the three subjects need to obtain their weights. We adopt two methods: Hysical questionnaire survey and online survey, and distribute them together with the teaching quality evaluation indicators to all respondents, including students, teachers, and supervisors. Not only does it cover various populations, but it also ensures the objectivity of the survey. The survey table is shown in Table 3.4.

For the comparison of levels, the standard scaling method is used for quantification, and the specific scales are shown in Table 3.5.

2. Using the sum product method, normalize the factors in each column of the obtained judgment matrix, and the general term of the factors is Equation 3.5:

$$b_{ij} = b_{ij} / \sum b_{ij} (i, j = 1, 2, 3) \tag{3.5}$$

3. Add each normalized judgment matrix by row, and the general terms of the factors are shown in equation 3.6:

$$w_i = \sum_{j=1}^{3} b_{ij} (i = 1, 2, 3) \tag{3.6}$$

4. Normalize W again, as shown in equation 3.7:

$$w_j = w_i / \sum_{i=1}^{3} w_i (i = 1, 2, 3) \tag{3.7}$$

The obtained $W = (w_1, w_2, w_3)$ is the weight occupied by each subject.

However, the obtained weights need to be tested for reasonableness, which is called consistency testing. The calculation of consistency ratio is called C.R (Consistency Ratio), where CR=CI/RI. Among them, CI is called Consistency Index, which is used to determine consistency indicators, while RI is called Random Index, which is the average random consistency indicator. From $CI = (\lambda_{max} - n)/(n-1)$, CR can be obtained. If CR is less than 0.10, we determine that the result of the matrix is consistent with the actual situation. Finally, the M (A) obtained earlier is multiplied by W and summed to obtain the teaching quality value of our course teacher, which determines whether the course meets our desired results.

3.3. System Analysis.

3.3.1. System design objectives. This system aims to combine software engineering personalized teaching design with software engineering personalized teaching quality evaluation and improvement. By leveraging personalized indicators tailored to software engineering, a suite of personalized software tools is developed to cater to individual student preferences and learning styles. These tools are utilized to evaluate the teaching quality of course instructors in software engineering, particularly in personalized teaching contexts. This initiative is instrumental in assessing and enhancing the teaching quality within Problem-Based Learning (PBL) modes, thereby driving advancements in teaching methodologies within the School of Software at Nanchang University [17]. In summary, the design objectives of the system are as follows:

1. The design of the system has practicality, applicability, and reliability.
2. The design of the database is logical and scalable.
3. The extraction and read in operations between data have a certain degree of flexibility and will not cause confusion.
4. Convenient for users to operate, reducing learning costs, and providing a simple and intuitive output interface.

3.3.2. System Feasibility Analysis. Nowadays, computers have become an essential part of people's lives, and network-based teaching quality evaluation is no longer limited to the collection of information through questionnaires in the classroom. Instead, students, teachers, and supervisors can conduct teaching quality evaluation anytime and anywhere, greatly improving efficiency. Questionnaire surveys have also become a supplementary way of data acquisition. In addition, the software engineering personalized teaching kit also meets the needs of student personality development. Students are also responsible for their knowledge when evaluating, and the school also improves the quality of their teaching, which is a win-win result. Economically speaking, it is feasible. In addition, by adopting the B/S structure mode and SQL Server 2005 database, the Software College can fully achieve self-sufficiency in terms of usability and operability, meeting the feasibility of operation and technical feasibility.

3.3.3. System database design. The system database uses SQL Server 2005 database management system. Below is a detailed explanation of the creation and logical structure of each data table in the database.

1. Overall design of E-R diagram

The overall logic design of E-R is shown in Figure 3.1.

2. Design of overall Table

Student Table: Including student name, student ID, gender, password, personalized indicators, and other data items, with the primary key being the student ID;

Teacher Table: Including teacher name, teacher ID, professional title, password, etc., with the primary key being the teacher ID;

Student transcript: Including student ID, course ID, grades, etc;

Supervisor Table: Including supervisor name, job number, password, etc., with the primary key being the supervisor employee number;

Administrator Table: Including administrator ID and password, with the primary key being administrator ID;

Curriculum schedule: Including course ID, teacher ID, course name, course credits, etc., with the primary key being course ID;

Teaching quality evaluation form: Including teaching ID, start time, teacher ID, course ID, evaluation score, etc. The primary key is the teaching ID (the same teacher, same course, courses opened at different

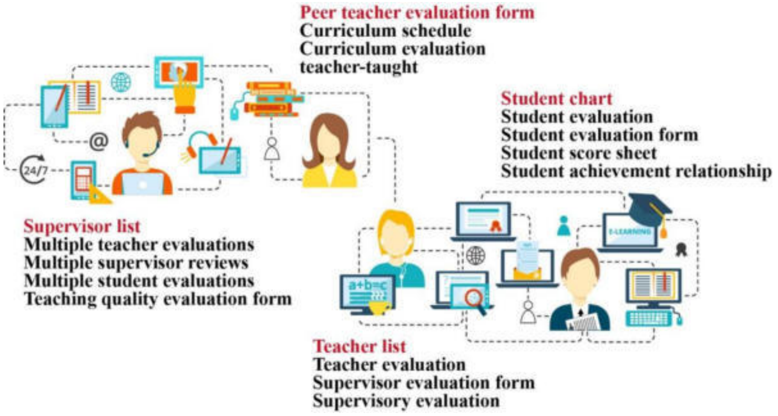


Fig. 3.1: E-R Logic Structure Diagram

times have a unique teaching ID representation, because the same teacher can take the same course in different years, it must be distinguished);

Student evaluation form: Including student ID, student evaluation indicators, and teaching number, with the primary key set to auto increment;

Peer teacher evaluation form: Including teacher ID, peer teacher evaluation indicators, and teaching number, with the primary key set to auto increment;

Supervisor evaluation form: Including supervisor ID, supervisor evaluation indicators, and teaching number, etc., with the main key set to auto increment.

4. Result analysis. After solving how to determine personalized courses for students, it is necessary for this group of students to conduct quality evaluation of software engineering personalized teaching. Because the students who participated in the evaluation were identified, the author will select evaluation information from students who are suitable for teaching quality evaluation. The data will be collected through a questionnaire on student teaching quality evaluation indicators, as well as 10 evaluations from peer teachers and 1 evaluation from supervisors.

4.1. Simulation method for teaching quality evaluation data. Prior to delving into the analysis of teaching quality evaluation, as previously discussed, it is crucial to assess the three primary stakeholders involved and assign appropriate weights to their evaluations. This entails determining the significance of each subject's assessment. Subsequently, utilizing their respective evaluation criteria, we compute the score for each subject's evaluation. These scores are then multiplied by the weights derived from the Analytic Hierarchy Process (AHP). Ultimately, this process yields the course's teaching quality level score, providing a comprehensive measure of teaching effectiveness[18].

4.2. Compilation of Teaching Quality Evaluation Data.

1) Calculation of weights for the three major entities. There are three types of subjects for evaluating teaching quality. One type is supervision, which comprehensively evaluates the teaching quality of the course through listening to lectures, represented by T1; Another type is peer teachers who discover the comprehensive quality of teaching by observing courses, represented by T2; The last category is students, mainly evaluating their own satisfaction and interest in the course, represented by T3. Summarize the results of all survey questionnaires and calculate the average value. The average value that T3 is considered more important than T2 is 3.6431; The average number of people who believe that T3 is more important than T1 is 2.7794; The average number of people who believe that T2 is more important than T1 is 1.4006. Based on the above data, a judgment matrix can be established as shown in Table 4.1.

After normalization, the matrix in Table 4.2 is obtained. By obtaining the average values of each row from the matrix, it can be calculated that T3=0.609498, T2=0.208339, T1=0.182164. After calculating the results, it

Table 4.1: Judgment Matrix

	T3	T2	T1
T3	1	3.6320	2.7683
T2	1/3.6320	1	1.4005
T1	1/2.7683	1/1.4005	1

Table 4.2: Normalized Matrix

	T3	T2	T1	Sum of rows	The average of the sum of rows
T3	0.611774	0.680042	0.536444	1.828382	0.609387
T2	0.167846	0.186557	0.270281	0.625007	0.208228
T1	0.220048	0.133167	0.193043	0.546380	0.182053

Table 4.3: Multiplied matrices

	T3	T2	T1	Sum of rows	weight	Sum of rows/weight
T3	0.611774	0.680042	0.536444	1.828382	0.609387	3.075808
T2	0.167846	0.186557	0.270281	0.625007	0.208228	3.027541
T1	0.220048	0.133168	0.193043	0.546380	0.182053	3.020327

is necessary to check the results: Multiply the values obtained from T3, T2, T1 by each column in the original matrix to obtain the matrix shown in Table 4.3. The obtained λ_{max}=(3.075919+3.027652+3.020438)/3=3.041336. Consistency indicator C.I=0.020668. When N=3, the average random consistency index R.L=0.58, and the proportion of consistency is C.R = 0.020668/0.58 = 0.035634. The results demonstrate good consistency. The final weight coefficients between the three main entities are:

$$T1 = 0.182163, \qquad T2 = 0.208339, \qquad T3 = 0.609498$$

2) Calculation of teaching quality evaluation data. As mentioned earlier, $S =< U, R, V, f >$ and U represent the collection of all research subjects, students, teachers, and supervisors; R is the set of attributes of the subject, please refer to the appendix for specific attributes; V is a set of attribute values; Vr is the value of the value attribute value (1-5); F is an allusion. Because U has three main subjects, and its peer teachers have fewer scores and lower attribute values, it is easy to display. Therefore, the calculation process here takes the teacher subject as an example. Calculate $SigX(a_i)$ based on the attribute importance formula, and calculate $\lambda_i = Sig(a_i)/\sum_{i=1}^{n} sigx(a_i)$: $\lambda_i = (0.12, 0.12, 0.12, 0.12, 0.12, 0.09, 0.1, 0.06, 0.15)$ Determine the reliability $M(A_i)$ of each attribute a_i based on the information provided by the decision-maker, satisfying $M(A_i) = 1$.

$$M(A_i) = \{0.11, 0.20, 0.15, 0.17, 0.16, 0.11, 0.04, 0.03, 0.13\}$$

Then choose the appropriate empirical factor, as mentioned earlier, θ=[0,1]. A smaller value indicates a greater emphasis on objective weight, while a larger value indicates a greater emphasis on experience. Therefore, we choose θ=0.2 to place greater emphasis on objective weight. Calculate the comprehensive reliability $M_i'(A_i)$

$$M_i'(A_i) = \{0.118, 0.136, 0.126, 0.13, 0.128, 0.094, 0.088, 0.054, 0.146\}$$

Finally, based on the evidence theory synthesis formula, the score for each evaluation is obtained as shown in Figure 4.1.

Figure 4.1 shows the performance of file calculation processing.

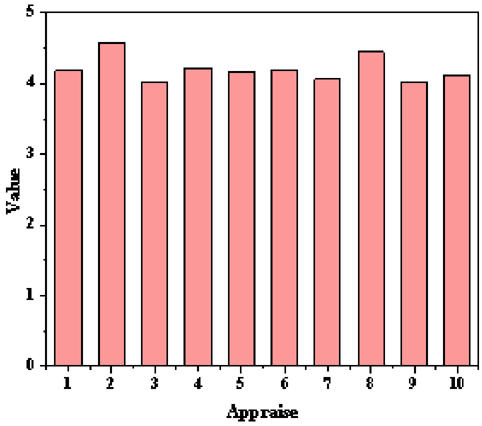


Fig. 4.1: Score for each evaluation

Finally, the evaluation of the webpage production engineering training course by peer teachers was 4.187 points. Using the same algorithm, the student evaluation score was 4.407 points, and the supervision score was 4.005 points.

Obtain a comprehensive score based on the weights of each entity obtained through the previous AHP Analytic Hierarchy Process

$$Q = 0.182183 * 4.016 + 0.208339 * 4.198 + 0.609498 * 4.418 = 4.2878$$

So the practical training score for webpage production engineering in this course is 4.2878 points.

4.3. Analysis of Teaching Quality Evaluation Data. The advantage of using the Analytic Hierarchy Process (AHP) lies in overcoming the influence of human bias and expectations on the allocation of weights between subjects. Through the method of group questionnaire surveys and an organic combination of experience and mathematical methods, it achieves objectivity, rationality, impartiality, scientificity, and persuasiveness. From the weight of the results, it can also be clearly reflected that personalized teaching in software engineering is a student-centered basic concept. In the rough evidence set algorithm, $M(A_i)$ is obtained through experience. If $M(A_i)$ is simply used as the weight, the subjectivity is too heavy [19]. For example, a_2, a_3, a_4 obtains higher weights on $M(A_i)$. However, with the fusion operation of λ_i and the intervention of θ empirical factor, a balance point between subjectivity and objectivity is found. By combining the weights between subjects obtained through Analytic Hierarchy Process, the final score of 4.2989 is consistent with the survey and public opinion. This method avoids the deficiency of traditional evidence theory that treats all evidence equally, enhances the ability of information fusion, and obtains more realistic conclusions [20].

5. Conclusion. The author has established the weights of various evaluation indicators that affect teaching quality, calculated the evaluation weights among the three major subjects, established an initial data warehouse, and studied and established a personalized teaching quality evaluation model for software engineering, including design requirements, database documents, and ER diagram use cases. Enable the model to be quickly transformed into a system and put into use. In this data acquisition and operation, based on the establishment of personalized courses and students, suitable courses were found and students were involved in the evaluation. Students who were not suitable for the evaluation were eliminated, and an improved algorithm combining AHP Analytic Hierarchy Process and Rough Evidence Set Theory was used to successfully obtain the conclusions we need, further verifying the feasibility and usability of the software engineering personalized teaching quality evaluation and improvement model.

Acknowledgement. 2023 Research Project on Educational Teaching Reform in Hainan Higher Education Institutions, Research on the Application of Big Data in Vocational Undergraduate Finance, Economics and Trade Majors in the Age of Digital Intelligence, Hnjg2023ZD-62.

Key Laboratory of Philosophy and social Science in Hainan Province of Hainan Free Trade Port International Shipping Development and Property Digitization, Hainan Vocational University of Science and Technology Hainan Social Science [2022] No. 26

2023 Hainan Higher Education Teaching Reform Research Fund Project Hnjg2023-148: Research on Professional Setting and Dynamic Adjustment Mechanism of Higher Vocational Colleges and Universities under the Perspective of Industrial Demand.

REFERENCES

[1] Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and information technologies, 26(1), 205-240.

[2] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 10(3), 1355.

[3] Martínez-Abad, F., Gamazo, A., & Rodriguez-Conde, M. J. (2020). Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. Studies in Educational Evaluation, 66, 100875.

[4] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. Knowledge-Based Systems, 200, 105992.

[5] Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), 237.

[6] Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 143, 103676.

[7] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 11.

[8] Yang, Y. H. . (2022). Research on the teaching quality management of"xindi applied piano pedagogy". Contemporary Education Studies (Photo), 6(7), 80-87.

[9] Pertuz, S. , Ramirez, A. , & Reyes, O. M. . (2022). Course quality assessment in post-pandemic higher education. 2022 IEEE Learning with MOOCS (LWMOOCS), 120-125.

[10] Che, Y. , Che, K. , & Li, Q. . (2022). Application of decision tree in pe teaching analysis and management under the background of big data. Computational intelligence and neuroscience, 17(1), 3.

[11] Zhao, L. . (2022). Interactive teaching mode based on big data in blended english teaching. 2022 International Conference on Education, Network and Information Technology (ICENIT), 265-268.

[12] Gamazo, A., & Martínez-Abad, F. (2020). An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. Frontiers in Psychology, 11, 575167.

[13] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

[14] Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. Expert Systems with Applications, 166, 114060.

[15] Kumar, T. S. (2020). Data mining based marketing decision support system using hybrid machine learning algorithm. Journal of Artificial Intelligence, 2(03), 185-193.

[16] Ahirwar, M. K., Shukla, P. K., & Singhai, R. (2021). CBO-IE: a data mining approach for healthcare IoT dataset using chaotic biogeography-based optimization and information entropy. Scientific Programming, 2021, 1-14.

[17] Pan, Y., & Zhang, L. (2021). A BIM-data mining integrated digital twin framework for advanced project management. Automation in Construction, 124, 103564.

[18] Alzahrani, B., Bahaitham, H., Andejany, M., & Elshennawy, A. (2021). How ready is higher education for quality 4.0 transformation according to the LNS research framework?. Sustainability, 13(9), 5169.

[19] Raffaghelli, J. E., Manca, S., Stewart, B., Prinsloo, P., & Sangrà, A. (2020). Supporting the development of critical data literacies in higher education: Building blocks for fair data cultures in society. International Journal of Educational Technology in Higher Education, 17, 1-22.

[20] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130-160.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Apr 5, 2024
Accepted: May 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 231–240, DOI 10.12694/scpe.v26i1.3744

AN INTELLIGENT MONITORING SYSTEM FOR SPORTS MENTAL HEALTH STATUS BASED ON BIG DATA

YOULIANG HAN[*] AND WENGUANG GENG[†]

Abstract. This provides an effective way to break away from traditional inefficient evaluation methods for monitoring and analyzing the psychological status of a large number of school sports athletes, the author proposes an intelligent monitoring system for sports psychological health status based on big data. Applying big data technology to mental health assessment, using real-time monitoring and analysis of athlete unified EEG waves, dividing athlete EEG waves into frequency bands, and conducting mental health analysis. The author validated the effectiveness of the system through simulation experiments, and the results showed that the psychological states of the subjects were not the same during the early and fatigue stages of training. In the early stages of training, the brainwave frequency band was mainly in the Beta and Gamma bands, accounting for 37% and 41%, respectively. Concentration was greater than relaxation, while in the fatigue stage of homework, the brainwave frequency band was mainly in the Delta and Thata bands, accounting for 43% and 45%, respectively, and concentration was less than relaxation. The psychological monitoring system designed by the author can provide a technical foundation for a series of strategies to promote training efficiency while ensuring the mental health of athletes.

Key words: Big data, Psychological health, EEG signals, Real time monitoring

1. Introduction. In 2021, the General Office of the Ministry of Education issued a notice on strengthening the management of student mental health, emphasizing that mental health management should be strengthened in four aspects: source management, process management, result management, and guarantee management. In addition, the level of emphasis on student mental health education varies among different stages of education [1]. However, many families have significant gaps in mental health education, and vocational and secondary vocational colleges have limitations in understanding the mental health of students.

In order to further improve the pertinence and effectiveness of student mental health work, the General Office of the Ministry of Education issued a notice in 2021 on strengthening student mental health management work (Education and Political Affairs Office Letter [2021] No. 10), and proposed strategies such as ”strengthening process management” and ”early classification and relief of various pressures” to strengthen professional support and scientific management, and improve student mental health literacy.

According to statistics, about 10% of adolescents in China require mental health interventions. However, the development of psychological counseling is still in its early stages, and the market has shown an explosive growth trend. However, campus psychological counseling services still lack professionalism and systematicity [2]. Despite the establishment of various psychological counseling rooms, the campus is still a place with a high incidence of depression and other mental illnesses, and has not achieved the expected results [3]. With the advent of the big data era, the data reserves and technological concepts related to big data are predicting the development trends of things in an unprecedented way, changing the knowledge system, lifestyle, and mental health level of students [4]. At present, the psychological construction work of schools is only limited to hiring a small number of professional psychological counseling teachers, while class teachers (non psychology professionals) mainly undertake manual intervention and management of students' psychological situations. Sports mental health plays a crucial role in the performance and overall health of athletes. With the development of big data technology, researchers have begun to explore how to use big data technology to monitor and evaluate the mental health status of athletes. The research on this big data based intelligent monitoring system for sports mental health status aims to combine advanced data collection techniques, data analysis methods, and psychological principles to provide comprehensive and timely psychological health monitoring

[*] Jiangsu Maritime Institute, Sports department, Nanjing, 211170, China (Corresponding author, 13913015815@163.com)

[†] Department of Physical Education, Nanjing Agricultural University, Nanjing, 210095, China

and intervention support for athletes. In traditional sports mental health monitoring, athletes usually rely on self-report or professional psychological assessment tools, which has problems such as strong subjectivity and untimely information acquisition. A monitoring system based on big data can comprehensively and objectively understand the psychological state of athletes through the collection and analysis of multi-source data such as daily training data, competition data, physiological parameter data, and social media data. This system can identify the psychological health status and changing trends of athletes, such as anxiety, stress, confidence, etc., by analyzing a large amount of data. At the same time, by combining with the personal characteristics and historical data of athletes, personalized psychological health intervention suggestions can be provided to help athletes better cope with challenges and stress.

Overall, the research on an intelligent monitoring system for sports mental health status based on big data can not only provide athletes with more comprehensive and objective mental health monitoring services, but also provide personalized psychological support and intervention, thereby improving their competitive performance and overall health level.

2. Literature Review. Wang, K. et al. advocate for early intervention in mental health disorders, particularly among adolescents and children. They emphasize the importance of establishing various positive psychological education frameworks for individuals grappling with mental health issues. Consequently, the development of effective mental health education leveraging big data is crucial to fostering positive thinking and support systems, especially within school environments.Ultimately, this innovative mental health education model enhances individuals' learning experiences and contributes to a reduction in mental illness, depression, and stress within communities[5]. Wang, S. et al. introduced an innovative empirical framework that harnesses the power of social media to systematically evaluate, quantify, map, and monitor a nation's mental health status. This framework is structured to enable ongoing surveillance and scalability to other countries as needed. By tracking regions where individuals express heightened levels of negative mental health indicators via social media, valuable insights are gleaned to strategically allocate limited mental health resources. This approach facilitates intelligent resource allocation, ensuring targeted interventions where they are most needed based on real-time data analysis of digital expressions of mental well-being[6]. Liu, X. et al. conducted a comprehensive study involving two rounds of data collection and organization, followed by rigorous statistical analyses such as descriptive analysis, unbiased t-tests, chi-square tests, variance assessments, and SNK-q tests using validated data. They investigated the evolution of mental health among students experiencing negative psychological symptoms over a two-year period and assessed the associated impacts. Employing advanced machine learning techniques, they developed models to identify susceptibility factors and analyze their contribution to mental health outcomes. The findings from testing and data analysis confirm the efficacy and viability of their approach[7]. Bakare, A. et al. explored the processes involved in gathering, monitoring, managing prescriptions, and analyzing real-time health data. They employed a network simulator to evaluate the effectiveness of their proposed system and compared it with various communication protocols to identify optimal techniques for health monitoring[8].

By applying big data technology to the monitoring and analysis of the psychological health of sports students, corresponding training allocation plans can be formulated based on work characteristics and the psychological status of sports students. Real time tracking of the psychological status of sports students can be carried out during the training process, and intervention can be given at appropriate time points, thereby achieving the effect of improving the experience of sports students while ensuring efficiency. Therefore, by combining big data monitoring and analysis technology with the theory of EEG psychological feedback for sports students, a psychological health evaluation system for real-time monitoring and analysis of the psychological state of sports students is designed, providing a new practical path for related fields.

3. Research Methods.

3.1. Establishment of a data-driven system for brainwave monitoring and analysis . The author designs an automatic monitoring system for psychological health evaluation data based on big data perception technology, which collects different EEG signals using big data perception technology to automatically monitor and analyze human psychological health [9]. The author mainly understands the essence of mental health from two aspects. On the one hand, it evaluates the subjective state and direct indicators of work and learning ability. On the other hand, it analyzes the indirect indicators of work and learning ability. Compared with direct indicators, indirect indicators can more accurately reflect the psychological health of testers. The method used needs to monitor the initial symptoms of the tester when their mental health is relatively good. For fluctuations in their mental health that conform to normal conditions, an appropriate sensitivity range should be set for the monitoring indicators.

The electrical signals of the brain are generated by the stimulation of ion movement by neurons, and all reactions and actions in the human body are completed by the potential changes generated by synapses between countless neurons in the brain [10]. EEG changes can be divided into evoked potential response and self generating activity. Spontaneous EEG is similar to sinusoidal signals, and it changes over time without specific external stimuli. Due to the fact that the time-domain waveform of spontaneous EEG waves does not have specific patterns, but it conforms to specific patterns in the frequency domain, it is generally classified based on the frequency domain of EEG waves. EEG waves are divided into five different rhythms based on frequency, including Y waves (31 Hz to 100 Hz), β waves (14 Hz to 30 Hz), α waves (8 Hz to 13 Hz), θ waves (4 Hz to 7 Hz), δ waves (1 Hz to 3 Hz). The voltage amplitude value of Y wave is detected as $1\ \mu V \sim 5\mu V$. Indicates that the tester is currently highly mentally tense and generates EEG pulses when stimulated, with intermittent buffering between them [11]. The voltage amplitude value of β waves is detected as 5 uV-20 uV, with a high frequency, indicating that the tester is currently relatively nervous and has a high level of perception of the surrounding things. The voltage amplitude value of waves is detected as 20uV-100 uV, and its frequency is relatively stable in the absence of external stimuli, indicating that the tester's brain is currently clear and relaxed, with more focused attention, and in a brain state suitable for work and learning.

The voltage amplitude value of θ waves is $50\ \mu V \sim 150\mu V$. The amplitude of EEG waves is relatively stable, indicating that the tester's current mental state is relatively relaxed and their attention concentration is poor, gradually entering a state of fatigue. When exposed to external stimuli, their attention will be focused. The voltage amplitude value of δ wave is $20\ \mu V \sim 200\mu V$. Indicates that the tester is currently in a state of extreme fatigue. When the tester gradually wakes up due to external stimuli, there may also be discontinuities δ wave, θ wave sum δ wave.

In the actual detection of EEG signals, the collected signals are presented as superimposed time-domain waveforms of EEG signals with different rhythms, making it difficult to obtain EEG information that can describe the psychological health of the tester. Therefore, it is necessary to convert the time-domain signals into frequency-domain signals and analyze them. Calculate using the eSense index based on the proportion of brain wave energy with different rhythms obtained [12]. This index is mainly used to describe the focus and relaxation of testers in reflecting their mental health status during the work and learning process. Y wave and β when the proportion of waves in the energy of EEG signals is relatively high, it indicates that the tester is focused and mentally tense; θ wave sum δ when the proportion of waves in the energy of EEG signals is high, it indicates that the tester has poor concentration and relatively relaxed mental state. The formula for calculating focus is shown in equation 3.1.

$$P_1 = (mY + n\beta + t\alpha) \times 100 \tag{3.1}$$

In equation 3.1, Y,β and α represent Y waves, respectively, β wave sum α proportion of waves in the energy of EEG signals, where m, n, and t represent Y waves, respectively, β weight coefficients of wave and α wave. The calculation formula for relaxation is shown in equation 3.2.

$$P_2 = (x\theta + y\delta + z\alpha) \times 100 \tag{3.2}$$

In equation 3.2, θ δ and α represent separately θ wave δ wave sum α proportion of waves in the energy of EEG signals, represented by x, y, and z, respectively θ wave, δ wave sum α weight coefficient of the wave. Set a rating range of 1-100 for the focus and relaxation of the tester, and evaluate their current thinking and mental state [13,14].

3.1.1. Design of Big Data Brain Wave Analysis System. When designing an automatic monitoring system for mental health assessment data, the author mainly divided the system into two parts, the first part being the big data monitoring part and the second part being the big data analysis part. The big data monitoring

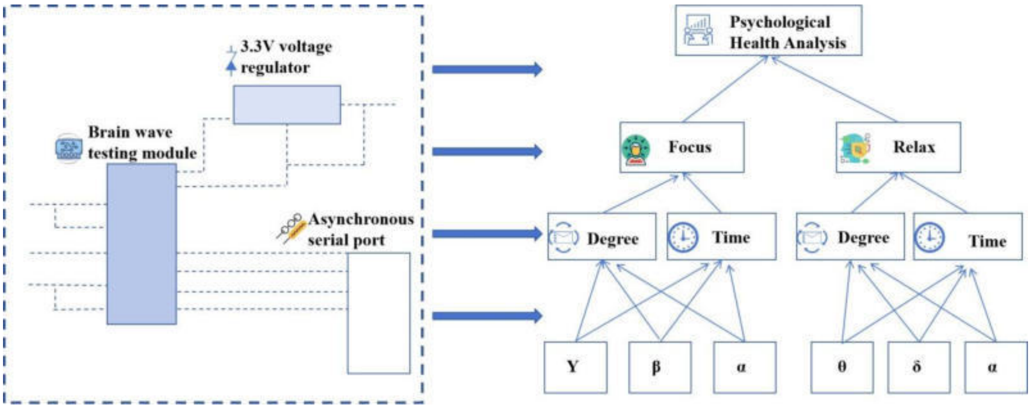


Fig. 3.1: Architecture of Big Data Brain Wave Monitoring and Analysis System

part mainly adopts cluster based EEG monitoring as the main monitoring method, and the uploading of EEG signals is also in the form of cluster uploading. The big data analysis part is responsible for cluster analysis of EEG data and ultimately forming psychological health analysis results. The system structure is shown in Figure 3.1.

From Figure 3.1, it can be seen that the monitoring part of the system is mainly composed of a brainwave testing module, equipped with 3 A 3V voltage regulator and asynchronous serial port, this module can complete functions such as brain wave signal acquisition, signal filtering, signal scaling, and signal conversion [15]. In the process of information collection, subjects do not need to undergo traumatic monitoring, but only need to use a head mounted contact point, which is more suitable for the scenario of work psychological monitoring.

In the big data analysis section, combined with the established brain wave monitoring and analysis data-driven system, the system strictly analyzes the changes in different bands of brain waves under external stimuli through big data hierarchy analysis, transforming brain wave performance into two dimensions of psychological focus and psychological relaxation, among them, the bottom level brain wave judgment of psychological focus is based on Y waves, β wave sum α three types of EEG signals are used as the basis, while the bottom level EEG judgment of psychological relaxation is based on θ wave, δ wave sum based on three types of EEG signals. In the scenario of considering work efficiency, the ratio of worker's focus to focus time and the ratio of relaxation to relaxation time is 2, and the matrix is shown in equation 3.3.

$$A_a = \begin{bmatrix} 1 & 2 \\ 1/2 & 1 \end{bmatrix} \tag{3.3}$$

The focus judgment matrix obtained from the analysis is shown in equation 3.4.

$$B_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1/2 & 1 & 3/2 \\ 1/3 & 2/3 & 1 \end{pmatrix} \tag{3.4}$$

The focus duration judgment matrix is shown in equation 3.5.

$$B_2 = \begin{pmatrix} 1 & 1/3 & 1/2 \\ 3 & 1 & 3/2 \\ 2 & 3/2 & 1 \end{pmatrix} \tag{3.5}$$

The relaxation judgment matrix is shown in equation 3.6.

$$C_1 = \begin{pmatrix} 1 & 1/4 & 1/2 \\ 4 & 1 & 2 \\ 2 & 1/2 & 1 \end{pmatrix} \tag{3.6}$$

The relaxation time judgment matrix is shown in equation 3.7.

$$C_2 = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 2 & 1 & 2/3 \\ 3 & 3/2 & 1 \end{pmatrix} \tag{3.7}$$

On this basis, the maximum eigenvalue of the judgment matrix is determined by judging the matrix, and consistency testing is carried out using the eigenvectors and the random consistency indicators found in the query. The results are represented according to the overall ranking, and the dual hierarchical overall ranking of psychological focus and psychological relaxation is finally obtained as shown in equation 3.8.

$$\begin{cases} W_1 = \begin{bmatrix} 0.89 \\ 0.74 \\ 0.50 \end{bmatrix} \\ W_2 = \begin{bmatrix} 0.32 \\ 1.01 \\ 0.75 \end{bmatrix} \end{cases} \tag{3.8}$$

In equation 3.8, W_1 represents the weight coefficient result of psychological concentration, and W_2 represents the weight coefficient result of psychological concentration. Through big data hierarchical analysis, it is possible to analyze the changes in brainwave rhythms of workers in batches, and to make automated judgments on the psychological changes of workers through a combination of quantitative and qualitative methods in a hierarchical and three-dimensional manner [16].

4. Result analysis. The research data comes from the fatigue assessment of sports athletes in a certain school in S Province. The author used wearable EEG detection together, the instrument used dry electrode collection form, and optoelectronic coupling isolation sensor technology to achieve noise reduction [17]. In terms of connection form, the study adopts a unipolar lead form, with the back of the tester's earlobe as the reference electrode. The fixed electrode is placed on the scalp, and the potential difference between the two is recorded. The research designed system can conduct unified real-time testing and big data perception analysis on the psychological concentration and relaxation of subjects. Through two-dimensional decomposition analysis of the psychological state of subjects, dynamic real-time warning can be achieved when the psychological concentration or relaxation of subjects reaches the warning value during the work process, thereby avoiding unnecessary errors caused by psychological problems during the training process. Before conducting actual experimental analysis, the author first conducted brainwave testing analysis on the system, and the specific analysis results are shown in Figures 4.1 and 4.2.

From Figures 4.1 and 4.2, it can be seen that the author divides the different rhythmic segments of EEG waves into different frequency bands from low to high, namely Delta waves (0 Hz to 3 Hz), Thata waves (4 Hz to 7 Hz), Low alpha waves (8 Hz to 10 Hz), High alpha waves (11 Hz to 13 Hz), Low beta waves (14 Hz to 22 Hz), High beta waves (22 Hz to 30 Hz), Low gamma waves (31 Hz to 46 Hz), and Midgamma waves (above 46 Hz).

Different bands display different wavelengths and rhythmic energy amplitudes in the same interval, indicating that the system can clearly distinguish their overlapping and differential parts when facing different wavelengths. The special band peaks formed by Delta waves (0Hz∼3 Hz) and Thata waves (4 Hz∼7 Hz) are also perfectly restored. After conducting big data monitoring and input, analyze the concentration and relaxation status of the subjects [18]. It can be seen that although the noise generated by the subject's own body can have a significant impact on the system's band detection, it still has a good analytical effect on the patient's psychological focus or relaxation during the system detection process. Among them, the band of psychological relaxation is significantly lower than that of psychological attention, and the fluctuation range is larger. There is a clear intersection between the two. Although the noise band of the subjects is similar to the fluctuation area of the psychological attention band, it can be seen that the psychological attention band is not affected, and the energy proportion of different rhythm stages is successfully extracted. It can be seen from this that

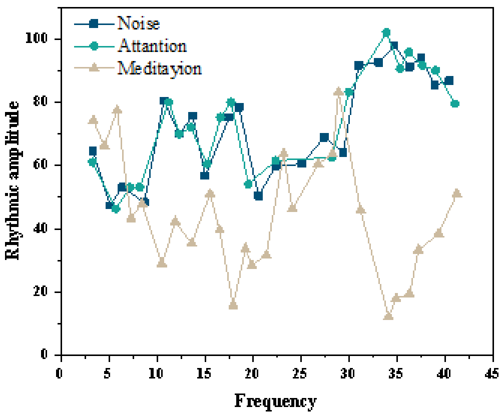


Fig. 4.1: Brain wave rhythm spectrum test chart

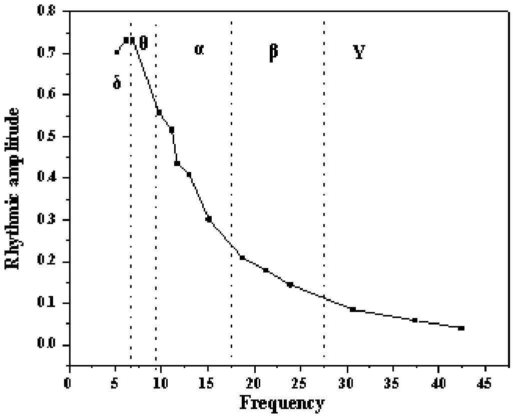


Fig. 4.2: Brain wave rhythm spectrum test rhythm spectrum

the psychological health evaluation data automatic monitoring system designed in the study is effective. The brainwave big data monitoring and analysis results of the subjects in the early stage of training are shown in Figures 4.3 and 4.4.

From Figures 4.3 and 4.4, it can be seen that the proportion of energy in the EEG rhythm spectrum of the subjects is relatively high in the Beta and Gamma bands of the subjects during the early training stage, among them, the energy proportion of the Beta band is 37%, while the energy proportion of the Gamma band is 41%. The Delta band and Thata band have the lowest proportion in the subjects' EEG waves, with the Delta band accounting for 8% and the Thata band accounting for 9%.

At the same time, observing the changes in the brainwaves of the subjects, it can be seen that the overall state of the subjects is relatively stable, and there are few cases of severe fluctuations, indicating that at the current stage, the subjects are in a highly focused state of attention. In this state, even under external stimuli, the subjects can still maintain their high concentration, that is, even if there are stimulating fluctuations in the brainwaves, they can quickly return to a normal and stable state. At the same time, from the comparison between the psychological concentration and psychological relaxation of the subjects, it can be seen that the psychological concentration line of the subjects is basically above the psychological relaxation line in the current situation, indicating that compared to psychological relaxation, the psychological concentration of the subjects is higher and maintains a relatively stable fluctuation range for a long time, showing a gradually decreasing

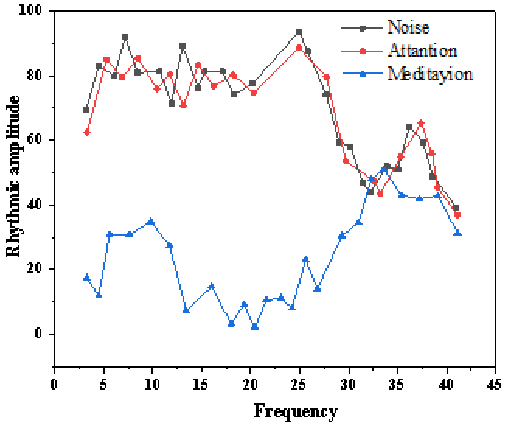


Fig. 4.3: Test chart of analysis results in the early stage of training

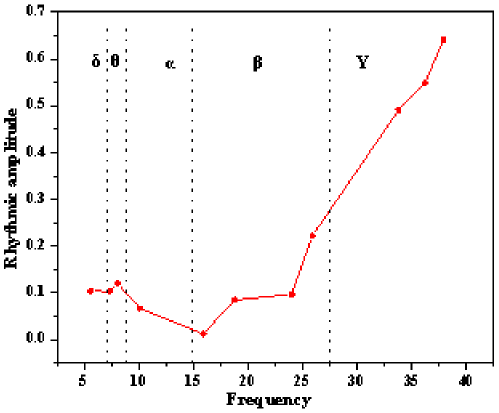


Fig. 4.4: Rhythm Spectrum of Analysis Results in the Early Training Stage

trend overall. On the other hand, the psychological relaxation line is opposite to the psychological concentration line, showing a gradually increasing trend. This indicates that with the increase of training time, the patient's psychological concentration gradually decreases, but the psychological relaxation gradually increases, and the two intersect in the later stage. From this, it can be seen that this stage is the stage in which the overall mental health status of the subjects is relatively good during the homework stage, and there is no need to intervene too much in the psychological status of the subjects [19,20]. But as the training time increases, the subjects will gradually shift from a more focused state to a more tired state during the training process, which is often the main interval for fluctuations in their mental health status.

The monitoring and analysis results of the subject's brainwave big data in this interval are shown in Figures 4.5 and 4.6.

From Figures 4.5 and 4.6, it can be seen that from the perspective of the proportion of energy in the EEG rhythm spectrum of the subjects, during the training fatigue stage, the energy proportion of the Delta and Thata bands in the EEG of the subjects is relatively large, with the Delta band accounting for 43% and the Thata band accounting for 45%. The Beta band and Gamma band have the lowest proportion of EEG waves in the subjects, with the Delta band accounting for 3% of energy and the Thata band accounting for 2% of energy. It can be seen that compared with the early stage of training, the proportion of brainwave energy in the subjects shows a completely opposite state, indicating that at this time, the subjects are generally

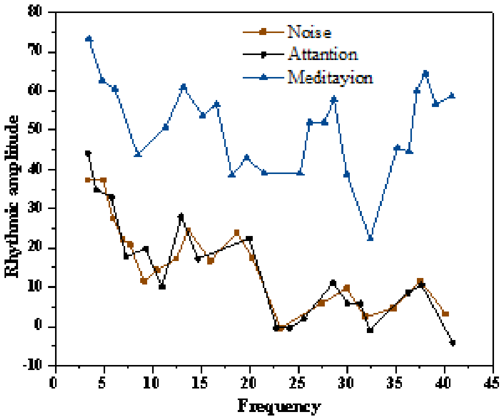


Fig. 4.5: Test chart of training fatigue stage analysis results

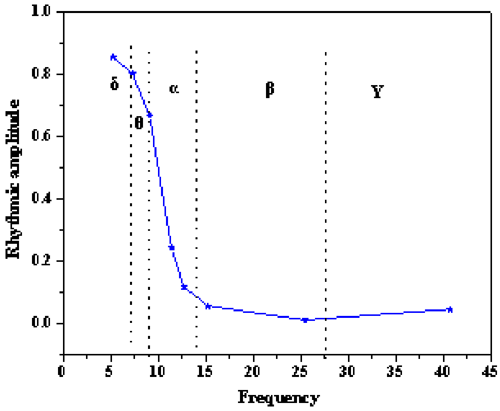


Fig. 4.6: Rhythm spectrum of training fatigue stage analysis results

in a state of fatigue and it is difficult to effectively concentrate their psychological attention. Observing the changes in the brainwaves of the subjects at the same time, it can be seen that compared to the early stage of training, the overall psychological state of the subjects at this stage is more unstable, prone to frequent and intense psychological fluctuations, and easily stimulated by the external environment. Moreover, once external stimuli are generated, the psychological state of the subjects is difficult to immediately recover, the frequency of being influenced by external stimuli is also constantly increasing. From the comparison between the psychological concentration and relaxation of the subjects, the psychological relaxation line is basically located above the psychological concentration line, indicating that compared to psychological concentration, the subjects have higher psychological relaxation, that is, their concentration is continuously decreasing and they are more relaxed. In this situation, the psychological concentration line of the subjects still shows a continuous downward trend, and the gap between their psychological concentration level and their level of psychological relaxation continues to widen. It indicates that the patient's psychological focus is generally in a declining state after the initial stage of training, but the level of psychological relaxation gradually increases in the early stage and stabilizes in the later stage. At this stage, the subjects are relatively tired and require external psychological intervention [21,22].

5. Conclusion. In order to address the issue of real-time and dynamic psychological health assessment for school sports athletes during training, the author combines big data real-time monitoring and analysis technology with brain wave detection technology. Through batch and unified monitoring and big data analysis of the changes in athletes' brain waves during training, the author aims to understand the psychological health of the staff. The author used simulation experiments to test the effectiveness of the technology.

The research results show that in the early stage of training, the energy proportion of the Beta and Gamma bands in the brainwaves of the subjects is relatively large, accounting for 38% and 40% respectively. At this time, the subjects have a greater psychological focus than their psychological relaxation, and can still quickly recover under external stimuli. During the fatigue stage of training, the energy proportion of the Delta and Thata bands in the brainwaves of the subjects is relatively high, accounting for 44% and 46% respectively. At this time, the subjects have a lower psychological concentration than their psychological relaxation, making it difficult for them to recover their focused state under external stimuli and are extremely susceptible to external factors. From this, it can be seen that the athlete mental health detection system designed by the author can effectively detect and analyze the psychological health changes of athletes during the training process, helping athletes improve their training experience while ensuring their own training efficiency.

Acknowledgement. Construction and practical exploration of the trinity cooperative mode of ""teaching, training and competition"" in college physical education from the perspective of educational environment theory(2023JSJG283)

REFERENCES

[1] Hickey, B. A., Chalmers, T., Newton, P., Lin, C. T., Sibbritt, D., McLachlan, C. S., ... & Lal, S. (2021). Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. Sensors, 21(10), 3461.

[2] Hoover, S., & Bostic, J. (2021). Schools as a vital component of the child and adolescent mental health system. Psychiatric services, 72(1), 37-48.

[3] Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., ... & Jeste, D. V. (2021). Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 6(9), 856-864.

[4] Gautham, M. S., Gururaj, G., Varghese, M., Benegal, V., Rao, G. N., Kokane, A., ... & Shibukumar, T. M. (2020). The National Mental Health Survey of India (2016): Prevalence, socio-demographic correlates and treatment gap of mental morbidity. International Journal of Social Psychiatry, 66(4), 361-372.

[5] Wang, K. , Zhou, Y. E. , Xu, J. X. , & Yang, G. . (2022). Reviewing big data based mental health education process for promoting education system. Current Psychology(5), 41.

[6] Wang, S. , Huang, X. , Hu, T. , Zhang, M. , Li, Z. , & Ning, H. , et al. (2022). The times, they are a-changin': tracking shifts in mental health signals from early phase to later phase of the covid-19 pandemic in australia. BMJ global health, 6(S5), 9070-9084.

[7] Liu, X. . (2022). Optimization of college students' mental health education based on improved intelligent recognition model. Mathematical Problems in Engineering, 75(8), 1080.

[8] Bakare, A. , Dutte, N. , & Sanap, A. . (2022). Iot based intelligent healthcare monitoring system. 2022 IEEE Pune Section International Conference (PuneCon), 1-6.

[9] Kalisch, R., Köber, G., Binder, H., Ahrens, K. F., Basten, U., Chmitorz, A., ... & Engen, H. (2021). The frequent stressor and mental health monitoring-paradigm: a proposal for the operationalization and measurement of resilience and the identification of resilience processes in longitudinal observational studies. Frontiers in Psychology, 12, 710493.

[10] Logeshwaran, J., Malik, J. A., Adhikari, N., Joshi, S. S., & Bishnoi, P. (2022). IoT-TPMS: An innovation development of triangular patient monitoring system using medical internet of things. International Journal of Health Sciences, 6(S5), 9070-9084.

[11] Simpson, K. R. (2022). Maternal mental health. MCN: The American Journal of Maternal/Child Nursing, 47(1), 59.

[12] Graham, A. K., Lattie, E. G., Powell, B. J., Lyon, A. R., Smith, J. D., Schueller, S. M., ... & Mohr, D. C. (2020). Implementation strategies for digital mental health interventions in health care settings. American Psychologist, 75(8), 1080.

[13] Chew, A. M. K., Ong, R., Lei, H. H., Rajendram, M., KV, G., Verma, S. K., ... & Gunasekeran, D. V. (2020). Digital health solutions for mental health disorders during COVID-19. Frontiers in Psychiatry, 11, 582007.

[14] Weist, M. D., Hoover, S. A., Daly, B. P., Short, K. H., & Bruns, E. J. (2023). Propelling the global advancement of school mental health. Clinical Child and Family Psychology Review, 26(4), 851-864.

[15] Gaebel, W., Lehmann, I., Chisholm, D., Hinkov, H., Höschl, C., Kapócs, G., ... & Zielasek, J. (2021). Quality indicators for mental healthcare in the Danube region: results from a pilot feasibility study. European Archives of Psychiatry and Clinical Neuroscience, 271, 1017-1025.

[16] Kim, J., Cheon, S., & Lim, J. (2022). IoT-based unobtrusive physical activity monitoring system for predicting dementia. Ieee Access, 10, 26078-26089.

[17] Islam, M. M., Rahaman, A., & Islam, M. R. (2020). Development of smart healthcare monitoring system in IoT environment. SN computer science, 1, 1-11.

[18] Dhanushkodi, K., Sethuraman, R., Mariappan, P., & Govindarajan, A. (2023). An efficient cat hunting optimization-biased ReLU neural network for healthcare monitoring system. Wireless Networks, 29(8), 3349-3365.

[19] Wu, X., & Ma, X. (2021). RETRACTED ARTICLE: Data mining-based air pollution characteristics and real-time monitoring of college students' physical and mental health. Arabian Journal of Geosciences, 14(15), 1509.

[20] Bickman, L. (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. Administration and Policy in Mental Health and Mental Health Services Research, 47(5), 795-843.

[21] Vella, S. A., & Swann, C. (2021). Time for mental healthcare guidelines for recreational sports: A call to action. British Journal of Sports Medicine, 55(4), 184-185.

[22] Gouttebarge, V., Bindra, A., Blauwet, C., Campriani, N., Currie, A., Engebretsen, L., ... & Budgett, R. (2021). International Olympic Committee (IOC) sport mental health assessment tool 1 (SMHAT-1) and sport mental health recognition tool 1 (SMHRT-1): towards better support of athletes' mental health. British journal of sports medicine, 55(1), 30-37.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Apr 15, 2024
Accepted: May 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 241–249, DOI 10.12694/scpe.v26i1.3765

THE FACTORY SUPPLY CHAIN MANAGEMENT OPTIMIZATION MODEL BASED ON DIGITAL TWINS AND REINFORCEMENT LEARNING

XINBO ZHAO[*]AND ZHIHONG WANG[†]

Abstract. This paper introduces the ”digital twin” to solve the problem of material allocation and real-time scheduling in the warehouse site. This project intends first to establish mathematical modeling based on a digital twin unmanned warehouse and dynamically optimize materials in the unmanned warehouse by combining visual analysis and deep reinforcement learning. Then, a security sharing mechanism of digital twin-edge network data based on blockchain fragmentation is proposed. For twin models with time-varying characteristics, a multi-node adaptive resource optimization method such as multipoint cluster selection, local base station consistent access selection, spectrum and computational consistency is constructed. This is done to maximize blockchain business processing power. A two-layer near-end strategy optimization (PPO) algorithm is proposed to solve the adaptive resource optimization problem. Experiments have proved that this method can significantly improve the overall processing power of the blockchain. In addition, this method is more adaptable than conventional deep reinforcement learning.

Key words: Unmanned storage; Digital twins; Deep reinforcement learning; Dynamic scheduling optimization; Digital twin edge network; Blockchain sharding

1. Introduction. To achieve accurate scheduling and optimal allocation of resources, most of the existing methods use heuristic methods to convert multiple high-quality multi-objective programming problems into a single programming problem. Alternatively, vehicles, three-dimensional shelves, testing equipment, etc., are regarded as a resource, and the optimal decision method is adopted to solve the problem [1]. However, the existing methods are limited in computing power and flexibility and can not effectively deal with multi-frequency, uncertain quantity of goods arrival, shelf, AGV, forklift and other resource optimization allocation problems. This seriously affects the service level and efficiency of the warehouse system. As a frontier and hot spot in intelligent manufacturing and storage, a digital twin is introduced into unmanned storage in this paper. Literature [2] takes the digital twin five-dimensional model as an example to introduce the application of this model in the warehouse. However, this method is mainly used in the manufacturing industry and can only play a reference role. Literature [3] uses digital twins to develop a new multi-mode intelligent terminal to solve the problem that real-time interaction cannot occur in manufacturing. Literature [4] integrates cyber twins with digital twins to build a networked digital twin model and remote control system oriented to information-physical fusion. This paper takes ”digital twin” and ”unmanned storage” as the starting point to study the integration of ”multi-class resource scheduling” and ”efficient scheduling.” The working condition of the equipment is monitored in real-time utilizing the Internet and visualization to improve its working efficiency and accurate scheduling level. Therefore, the digital twin unmanned warehouse architecture with multi-level characteristics is constructed according to the characteristics of the unmanned warehouse operation process. A real-time map construction method of unmanned warehouses based on physical modeling and data service systems is proposed. Then, the resource scheduling problem of a digital twin unmanned warehouse based on deep reinforcement learning is studied utilizing multidimensional information fusion.

2. Digital twin unmanned storage system design.

2.1. System Architecture.

This paper establishes the architecture of a digital Twin unmanned warehouse (Figure 2.1 is referenced in Building the Digital Representation with Digital Twin using Microsoft stack).

[*] Liaoning University of International Business and Economics, Dalian Liaoning 116052, China

[†] Liaoning University of International Business and Economics, Dalian Liaoning 116052, China (Corresponding author, zhaoxinbo0124@ 163.com)

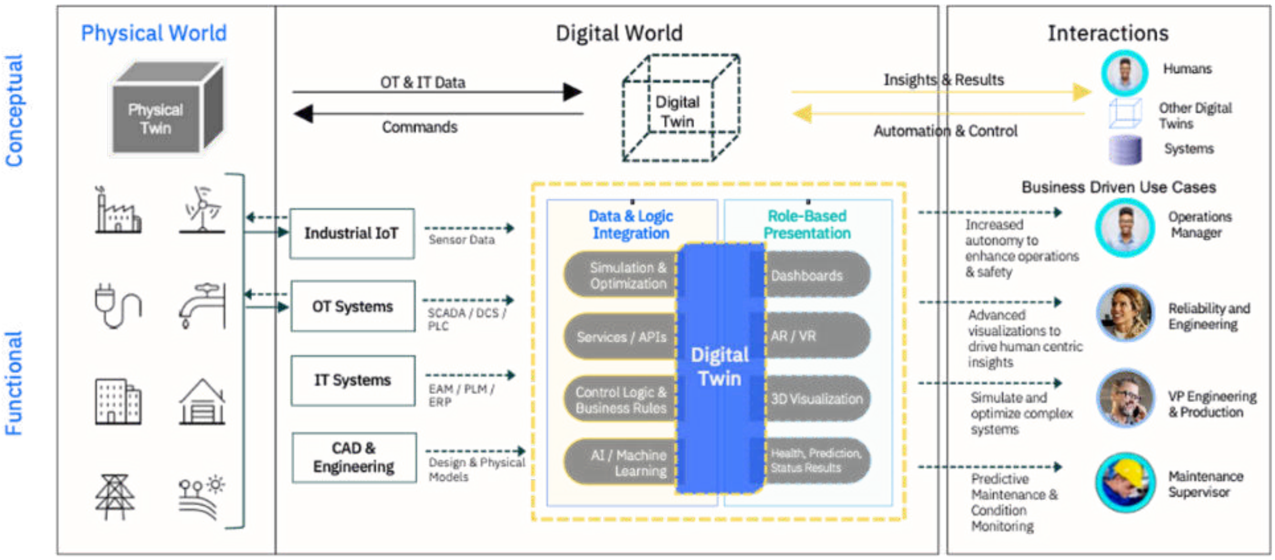


Fig. 2.1: Architecture of digital twin unmanned storage system.

The model consists of a visual, physical entity layer, two technology platforms, three layers, three databases, six logical process mappings, and six physical realities time mapping.

Among them, the sensing layer mainly identifies and acquires the target [5]. This layer uses the sensor node method to process the relevant data of shelves, forklifts, AGVs, goods, pallets, robots, warehouses, etc. and then transmits it to the corresponding location through relevant networking means to realize the collection of information on the lower layer.

At the data level, it realizes the management of user rights, the model interaction interface between the target model base, the real-time database and the local database [6]. Including warehouse signal, equipment status, equipment location, display information, warehouse location information, etc. Local databases include layout data, logical data, trigger mapping, initialization rules, scan point mapping tables, configuration files, etc. The system includes equipment data, production data, model base, operation base, order information, user personal information and so on.

At the business level, a predictive data-driven modeling method is adopted based on the twin data. This will make the warehouse management intelligent, thereby optimizing resource efficiency, optimizing the number of orders, optimizing the warehouse location, optimizing the area, and sharing resource information [7]. At the same time, the optimal results will also be transmitted back to the data center of the perception layer for virtual monitoring of the perception layer.

2.2. Elements of the operation process of digital twin unmanned warehouses. Among them, the operation process of a digital twin unmanned warehouse includes establishing the twin entity model, data system, and mapping logic.

2.2.1. Twin entity modeling. The ontology modeling method is used first when constructing twin entities. The ontology is constructed with class and attribute as the core. The category refers to the definition of the entity, and the property is the expression of the specific role of the class. The target and its properties must be modified before creation, and then its output is stored in the object library and recorded in the target table [8]. Lightweight methods are selectively adopted to reduce the display load during operation. The movable part in the 3D model is set as a movable body, and then the behavior trajectory of the movable part is modified. Animate it with associated components to form a whole. The elements of an unmanned warehouse and their relationship together constitute a complex network concept system. It includes 16 categories of objects, 21 connections, and 91 properties. You can see a detailed description in Figure 2.2.

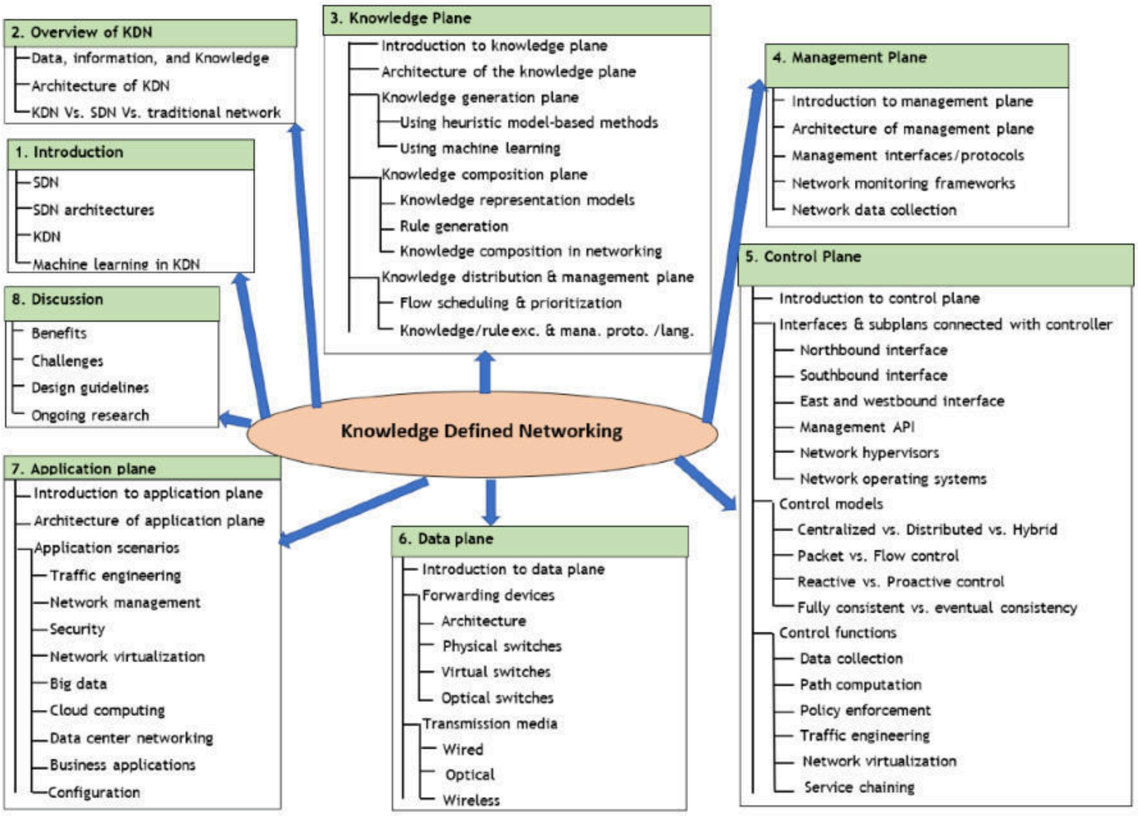

2. Overview of KDN
Data, information, and Knowledge
Architecture of KDN
KDN Vs. SDN Vs. traditional network
1. Introduction
SDN
SDN architectures
KDN
Machine learning in KDN
8. Discussion
Benefits
Challenges
Design guidelines
Ongoing research
3. Knowledge Plane
Introduction to knowledge plane
Architecture of the knowledge plane
Knowledge generation plane
Using heuristic model-based methods
Using machine learning
Knowledge composition plane
Knowledge representation models
Rule generation
Knowledge composition in networking
Knowledge distribution & management plane
Flow scheduling & prioritization
Knowledge/rule exc. & mana. proto. /lang.
4. Management Plane
Introduction to management plane
Architecture of management plane
Management interfaces/protocols
Network monitoring frameworks
Network data collection
Knowledge Defined Networking
5. Control Plane
Introduction to control plane
Interfaces & subplans connected with controller
Northbound interface
Southbound interface
East and westbound interface
Management API
Network hypervisors
Network operating systems
Control models
Centralized vs. Distributed vs. Hybrid
Packet vs. Flow control
Reactive vs. Proactive control
Fully consistent vs. eventual consistency
Control functions
Data collection
Path computation
Policy enforcement
Traffic engineering
Network virtualization
Service chaining
7. Application plane
Introduction to application plane
Architecture of application plane
Application scenarios
Traffic engineering
Network management
Security
Network virtualization
Big data
Cloud computing
Data center networking
Business applications
Configuration
6. Data plane
Introduction to data plane
Forwarding devices
Architecture
Physical switches
Virtual switches
Optical switches
Transmission media
Wired
Optical
Wireless

Fig. 2.2: Network structure of unmanned storage ontology model.

2.2.2. Data System Construction. Data service systems can realize real-time connections and calls between local and system databases. Run with a real-time database-driven model. A dynamic database-driven model is adopted. Entity knowledge ontology based on the OWL method is used to realize dynamic access to dynamic data in a database [9]. The database interface module is accessed regularly to realize the data analysis of the local database and system database. In this way, the unmanned storage environment can be quickly restored.

2.2.3. Mapping Logic. A geometric modeling method based on ontology is proposed, which can realize the unity of objects in space position, geometric size, motion characteristics, etc. The data service platform provides a unified control interface inside and outside the schema and interacts with the three databases [10]. Based on the law of real-time mapping, this paper efficiently combines the physical elements of the digital twin model to run the process of warehousing, tallying, storage, picking, order receiving, and delivery to the online process. This forms a complete unmanned warehouse business process. The logical flow of the real-time mapping is shown in Figure 2.3 (image cited in Developments in the Built Environment, Volume 17, March 2024, 100309).

2.3. Digital twin unattended warehouse scheduling optimization logic. The digital twin method is used to realize the effective utilization of the warehouse. Cluster analysis and deep reinforcement learning are used to analyze and optimize the resource effectiveness of the system [11]. After resource efficiency optimization, the allocation scheme is compared with that before optimization and fed back to the database for vector iteration to get the optimal solution. The specific content is shown in Figure 2.4.

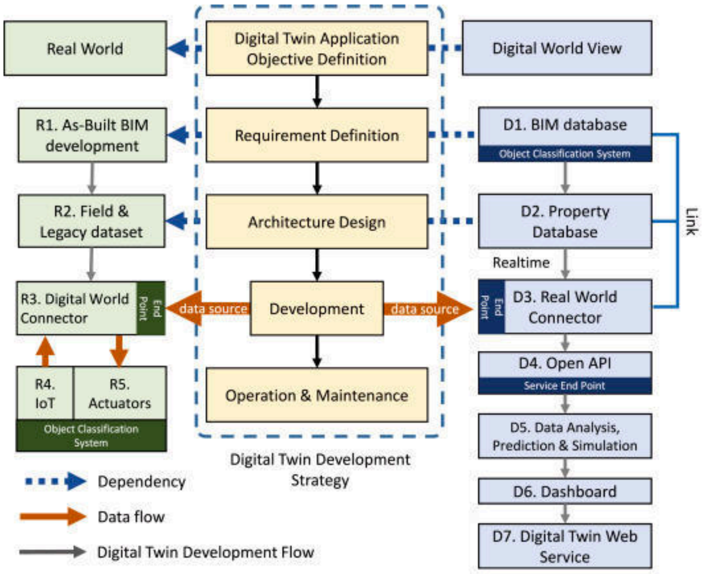

Real World
Digital Twin Application Objective Definition
Digital World View
R1. As-Built BIM development
Requirement Definition
D1. BIM database
Object Classification System
R2. Field & Legacy dataset
Architecture Design
D2. Property Database
Link
Realtime
R3. Digital World Connector
End Point
data source
Development
data source
End Point
D3. Real World Connector
R4. IoT
R5. Actuators
Object Classification System
Operation & Maintenance
D4. Open API
Service End Point
D5. Data Analysis, Prediction & Simulation
Digital Twin Development Strategy
D6. Dashboard
Dependency
Data flow
Digital Twin Development Flow
D7. Digital Twin Web Service

Fig. 2.3: Digital twin unmanned storage real-time mapping process logic diagram.

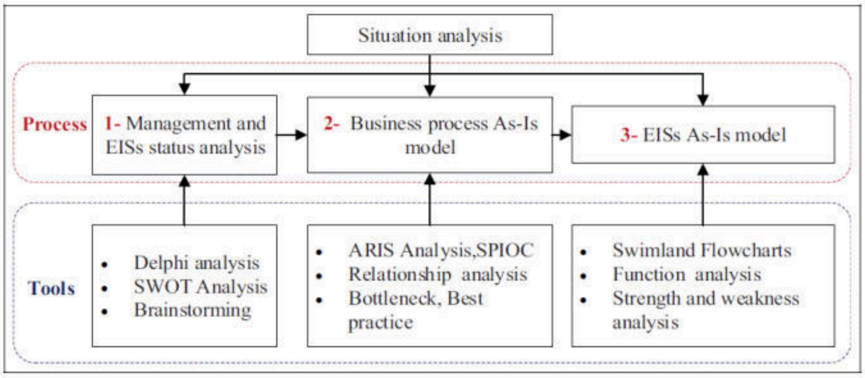


Fig. 2.4: Flow of resource efficiency optimization analysis.

2.3.1. Data analysis and prediction. The method of artificial neural network is adopted. Its input is based on the inbound and outbound commodity data collected by the digital twin data center, including the number of orders, order lines, received quantity, shipment quantity, inventory, dismantling amount, SKU and equipment status, etc., select the data related to the number of hidden layers, and divide it into training data, verification data and test data, the ratio of the three is about 7:1.5:1.5. The AUC value is used to determine the training effect, usually in the range of 0.5-1. The closer the value is to 1, the better the prediction effect of random judgment is. Combined with the collected data, the unmanned warehouse based on multidimensional information is scheduled, and its potential energy efficiency problems are fed back.

2.3.2. Automatic facility resource configuration. The automated system optimization process includes cluster analysis for device resource efficiency, and the generated unmanned warehouse must be encapsulated before it can be modeled to interact with deep reinforcement learning based on the Python language [12]. The unmanned warehouse model then takes the required form data from the cloud, executes the form and feeds

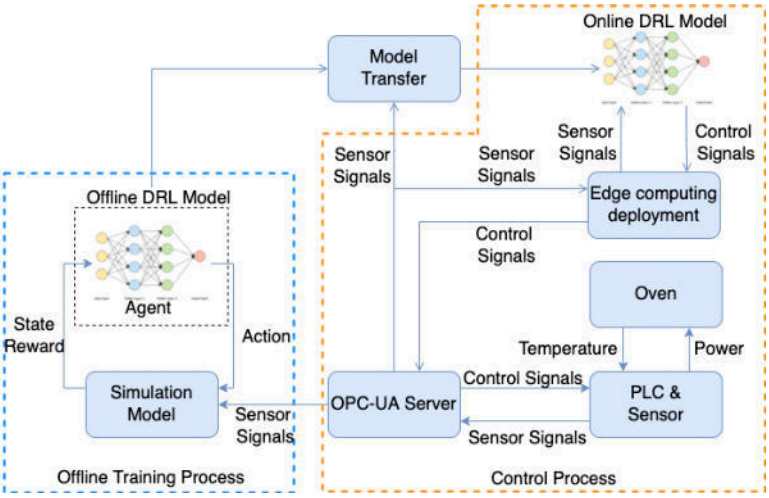


Fig. 2.5: Deep Reinforcement Learning Deployment Framework.

it back to the current state function. After obtaining the state matrix, the model makes decisions and operations. When the algorithm reaches the next determined time point, the algorithm will feed the current income and the state information of the next time point into the deep enhancement model. Finally, a trained deep reinforcement learning model for optimizing unmanned storage resources is obtained. The specific operation process is shown in Figure 2.5 (the picture is quoted in Using Deep Reinforcement Learning for Zero Defect Smart Forging).

2.3.3. Feedback of optimization results. Under the HTML architecture, the proven deep reinforcement learning mode is configured on Linux as an HTTP server. The jar bundle is configured in the cloud computing to access it as an API. In the implementation process, the data is collected, processed and integrated into the required data and uploaded to the cloud database [13]. In this paper, behavioral decisions are captured and fed back based on the deep reinforcement learning method of cloud computing and API models of the unmanned warehouse. Finally, the verified data is sent and returned to the terminal of the service layer. The user can see the optimal model and related parameters through the intuitive display interface. This makes the resource allocation of unmanned warehouses more scientific and reasonable.

3. Adaptive resource optimization method based on blockchain segmentation. The PPO method is a solid deep incentive learning method introduced by OpenAI in 2017, which is superior to other robust deep incentive learning methods in terms of sampling complexity [14]. By setting the trusted range, the method has an adaptive solid ability to avoid errors. Some scholars proposed adopting the PPO algorithm to adapt to the mapping error of digital twin model of the sheet metal assembly line to obtain the best clamping position. Currently, the edge data processing method based on the PPO method has a severe mapping error between the boundary twin and the physical network, and there are no corresponding research results. This paper studies the two-layer PPO algorithm to process different data types (Figure 3.1).

This paper presents a multi-layer PPO algorithm based on multiagent PPO. Block Administrator A uses a single PPO policy. Each K base station and block manager observed the existing dual-layer digital twin and imported the observations into the PPO neural network [15]. Finally, the boundary twin model was used to verify the output. Finally, the verified algorithm is optimized to the corresponding physical node. Compared with the conventional PPO method, this project proposes a dual multi-layer PPO method so that K BS and block administrators can obtain the data they need simultaneously, thus reducing the resource consumption required by manual intervention.

3.1. Application of two-layer PPO algorithm in digital twin. A two-level Markov decision model is constructed, and the model's state space, behavior space and reward function are studied.

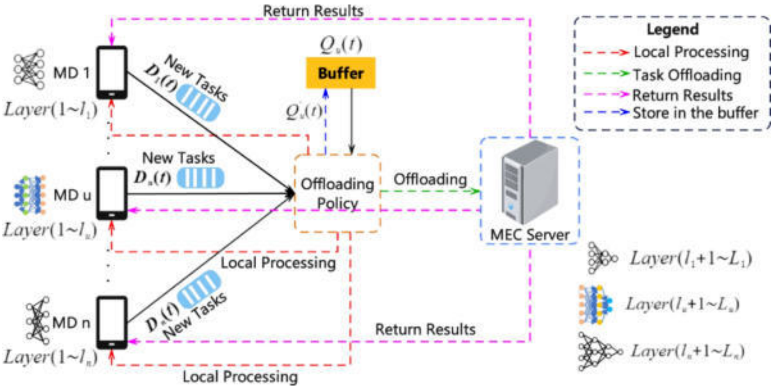


Fig. 3.1: PPO algorithm based on Downlink.

3.1.1. Phase space. At decision time $t(t = 1, 2, \ldots)$ there is a κ BS for maintaining the AP twin mode state of the local data sharing link. The algorithm includes the signal-to-noise ratio Θ_{n,j_k} of intelligent terminal n to $APjK$ and the signal-to-noise ratio $\Theta_{j\kappa,j\kappa'}$ of $APj\kappa$ to $APj\kappa'$, and $j_\kappa, j_\kappa{}' \in \mathfrak{J}_\kappa, j_\kappa \neq j_\kappa{}', n \in N_\kappa^\alpha, \alpha \in \alpha$. The state space of the κ base station is shown below

$$R_{\kappa,t}^z = \left[\Theta_{n,j_\kappa}, \Theta_{j_\kappa,j_{\kappa'}}\right]$$

K BS, whose state space is as follows

$$R_t^z = \left\{R_{1,t}^z, \cdots, R_{\kappa,t}^z, \cdots, R_{K,t}^z\right\}$$

In block administrator a, save the signal interference noise ratio $\Theta_{b,a}, \Theta_{a,\beta}, \Theta_{\beta,\beta'}, \Theta_{\beta,a}$ of the multicast transmission subchannel of the authentication node twin mode of the block, the maximum available computing resources $g_h^{\max}$ of the block manager and the authentication node, and $a, \beta, \beta' \in \beta, a \neq \beta, a \neq \beta', \beta \neq \beta'$. The block size is $R_{A,b}$, the local access is ASb and the block manager is a. Then, the state space of the block manager a can b expressed as:

$$| \; R_t^\beta = [R_{A,b}, \Theta_{b,a}, \Theta_{a,\beta}, \Theta_{\beta,\beta'}, g_h^{\max}]$$

3.1.2. Action space. The decision parameters of each node must be modified appropriately to meet the characteristics of time variability and maximize the benefit of K base stations and administrators of each layer [16]. The connection vector η between nodes and the bandwidth resource configuration vector Q_z of A-nodes are regulated in the local data sharing link. The local base station access vector is λ. The bandwidth resource configuration vector of the block manager and the parity node is Q_β. The resource allocation vector of the block manager and the parity node is g_β. In this way, the behavior space for optimal configuration of κ BS at the decision time t can be expressed as

$$H_{\kappa,t}^z = [\eta, Q_z]$$

Thus, the optimal behavior space for the utility of K BS can be expressed as

$$H_t^z = \left\{H_{1,t}^z, \cdots, H_{\kappa,t}^z, \cdots, H_{K,t}^z\right\}$$

In addition, the operation space for optimizing the utility of the regional manager a can be expressed as

$$H_t^\beta = [\lambda, Q_\beta, g_\beta]$$

3.1.3. Reward function. The constraints of C1 $-$ C7 must be verified in the operation of Layer 2PPO, so the following real-time reward function r_t is proposed in this paper. Here's $r_t^z = \sum_{\kappa \in K} K_{BS_x} \gamma_t^\beta = K_a^\beta$. If C1-C7 constraints cannot be met at the same time, it means that the current optimal strategy is not effective [17]. To prevent invalid decisions, the immediate return is set to 0 .

3.2. Dual-layer PPO algorithm principle. Combining deep neural networks with reinforcement learning solves the constructed two-layer Markov decision problem. The two-layer PPO method obtains the best-determining variable ξ^* by establishing the best parameters of the artificial neural network. This maximizes the average rate of return in formula

$$S(\xi) = \mathrm{E}_{\delta \sim \pi_{\xi}(\cdot r)} \left[\sum_{t=1}^{T} 1 r_t \right]$$

$0 \leq 1 \leq 1$ stands for discount factor. E stands for random sampling based on transformation order δ. The expected value of the immediate return is found given the strategy π_{ξ}, and the state $r.\delta$ represents the sequence of conditions and behavior changes at the corresponding time point t, which is $\delta = \{R_1^z, R_1^{\beta}, H_1^z, H_1^{\beta}, \cdots, R_t^z, R_t^{\beta}, H_t^z, H_t^{\beta}\}$. PPO is a reinforcement learning method that uses new strategy gradients and confidence intervals. The network of actors accepts the current situation of the actors and makes decisions accordingly [18]. The confidence interval method is used to dynamically adjust the parameters in the network so that the network has adaptive solid ability and good convergence. The loss function of the update process of the Actor-network parameter ξ_k^a of the κBS is expressed as

$$J\left(\xi_{\kappa}^{a}\right) = \min\left(\sigma_t\left(\xi_k^a\right) H_t, \operatorname{clip}\left(\sigma_t\left(\xi_{\kappa}^{a}\right), 1-\pi, 1+\pi\right) H_t\right)$$

$\sigma_t\left(\xi_k^a\right)$ indicates the updating range of network parameters. H_t represents the dominance function, which reflects the decision generated by the current network parameters. Compared to other possible decisions, $H_{k,t}^z$ is of superior value. $\pi \in (0,1)$ is the parameter that determines the upper and lower boundary $(1-\pi, 1+\pi)$ of the PPO algorithm's confidence range. clip$(\cdot)$ function is used to constrain $\sigma_t\left(\xi_k^a\right)$, so it has adaptive solid ability and convergence. The definition of $\sigma_t\left(\xi_k^a\right)$ is

$$\sigma_t\left(\xi_{\kappa}^{a}\right) := \frac{\pi_{\xi_k^a}\left(H_{\kappa,t}^z \mid R_{\kappa,t}^z\right)}{\pi_{\xi_{k,ald}^a}\left(H_{\kappa,t}^z \mid R_{\kappa,t}^z\right)}$$

ξ_{κ}^{a} represents the network parameters that have been updated. ξ_{κ}^{a},old is the network parameter before the upgrade. H_t is represented in formula (3.10). If the resulting decision $H_{\kappa,t}^z$ gets a better-expected return, then $H_t > 0$ is the opposite of $H_t < 0$. The dominant function H_t is defined in this way

$$H_t = \delta_t^z + \omega \delta_{t+1}^z + \cdots + \omega^{T-t+1} \delta_{T-1}^z$$

$\omega \in [0,1]$ stands for discount factor. δ_t^z represents the time error of a single step as defined below

$$\delta_t^z = r_{t+1}^z + \omega \rho\left(R_{t+1}^z\right) - \rho\left(R_t^z\right)$$

$\rho(\cdot)$ represents the Critic network's estimated reward for deciding $H_{\kappa,t}^z$. $r_{t+1}^z + \omega \rho\left(R_{t+1}^z\right)$ represents the sum of the immediate return r_{t+1}^z and the expected return of the Critic network corresponding to decision $H_{\kappa,t}^z$. The Critic neural network takes the change of the mean value of δ_t^z as its loss function. The weight ξ_k^z in the model is modified so that the cost function is maximum and the reward $\rho(\cdot)$ obtained by the algorithm is more accurate. The two different control strategies adopt the method of optimal learning rate to obtain the best network parameters $\xi_k^{a^*}$ and $\xi_k^{z^*}$.

4. System inspection. The project takes Z Company as an example to develop a digital twin unmanned warehouse system. According to the number of purchases made by the company between the first quarter of 2022 and the second quarter of 2023, they are classified and counted every week. The artificial neural network is used to analyze the actual production situation, and the AUC of 0.9245 is obtained, proving the method's effectiveness. This project adopts an A2C algorithm based on deep reinforcement learning for optimization [19]. The learning rate parameter was set to 1×10^{-6}, the simulation time step was set to 5min, the number of steps for each model training was 1000, and the total training step length was 5×10^{6}.

The tests compared to the inventory data are shown in Figure 4.1. Finally, the method is compared with those often used in unmanned warehouses. Higher rewards can be obtained through the optimal allocation

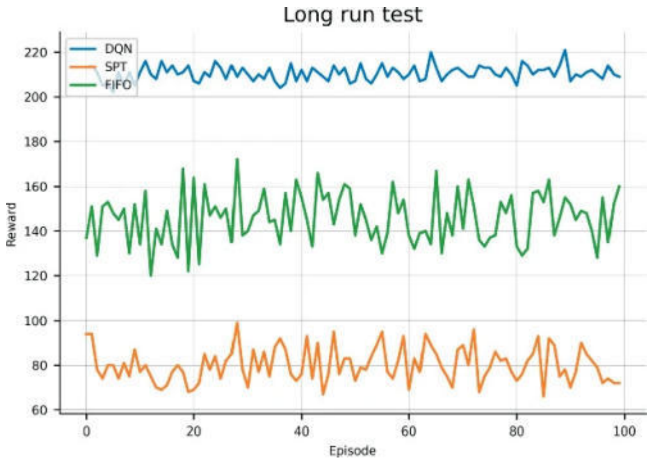


Fig. 4.1: Comparison of reward value, resource allocation and process time before and after optimization.

strategy. The deep enhancement method is adopted to optimize the design of forklift trucks in AGV, purchase area and loading and unloading area. The deep reinforcement learning method is used to configure the system resources dynamically, and the system's running speed is shortened from 26 minutes to 24.5 minutes. The time required to ship has been reduced from 3.6 points to 3.32 points. The material retention time in the warehouse was reduced from 44.21 minutes to 41.24 minutes. Through the dynamic resource adjustment of the system, the utilization rate and running speed are improved.

The best data information is returned to the data service system and is constantly adjusted to improve the model in the future. At the same time, these data will also be fed back to the data-sharing platform on the service side, and the decision maker can scientifically and reasonably allocate the corresponding resources according to the model and parameters on the visual interface.

5. Conclusion. A resource optimization method for unmanned warehouse systems based on deep reinforcement learning is proposed. This project uses simulation software to build a training environment for deep reinforcement learning. The model has effectively interacted with the production platform to realize the effective management of the authentic warehouse. An interactive model of an unmanned warehouse with man-machine interface is established. This project realizes collaborative optimization of unmanned warehouses based on cloud computing. This method has achieved good results in the actual operation of Z Company. The results prove the practicability of the proposed model, algorithm and prototype system.

Acknowledgement. The project is supported by the Scientific Research Project of Liaoning University of International Business and Economics(2023XJLXZD01).

REFERENCES

[1] Park, K. T., Jeon, S. W., & Noh, S. D. (2022). Digital twin application with horizontal coordination for reinforcement-learning-based production control in a re-entrant job shop. International Journal of Production Research, 60(7), 2151-2167.

[2] Bao, Q., Zheng, P., & Dai, S. (2024). A digital twin-driven dynamic path planning approach for multiple automatic guided vehicles based on deep reinforcement learning. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 238(4), 488-499.

[3] Yang, W., Xiang, W., Yang, Y., & Cheng, P. (2022). Optimizing federated learning with deep reinforcement learning for digital twin empowered industrial IoT. IEEE Transactions on Industrial Informatics, 19(2), 1884-1893.

[4] Shen, G., Lei, L., Li, Z., Cai, S., Zhang, L., Cao, P., & Liu, X. (2021). Deep reinforcement learning for flocking motion of multi-UAV systems: Learn from a digital twin. IEEE Internet of Things Journal, 9(13), 11141-11153.

[5] Badakhshan, E., & Ball, P. (2023). Applying digital twins for inventory and cash management in supply chains under physical and financial disruptions. International Journal of Production Research, 61(15), 5094-5116.

[6] Zhang, K., Cao, J., & Zhang, Y. (2021). Adaptive digital twin and multiagent deep reinforcement learning for vehicular edge computing and networks. IEEE Transactions on Industrial Informatics, 18(2), 1405-1413.

[7] Park, K. T., Son, Y. H., & Noh, S. D. (2021). The architectural framework of a cyber physical logistics system for digital-twin-based supply chain control. International Journal of Production Research, 59(19), 5721-5742.

[8] Goodwin, T., Xu, J., Celik, N., & Chen, C. H. (2024). Real-time digital twin-based optimization with predictive simulation learning. Journal of Simulation, 18(1), 47-64.

[9] Alexopoulos, K., Nikolakis, N., & Chryssolouris, G. (2020). Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing. International Journal of Computer Integrated Manufacturing, 33(5), 429-439.

[10] Lee, J., Azamfar, M., Singh, J., & Siahpour, S. (2020). Integration of digital twin and deep learning in cyber-physical systems: towards smart manufacturing. IET Collaborative Intelligent Manufacturing, 2(1), 34-36.

[11] Marmolejo-Saucedo, J. A. (2022). Digital twin framework for large-scale optimization problems in supply chains: a case of packing problem. Mobile Networks and Applications, 27(5), 2198-2214.

[12] Ren, Z., Wan, J., & Deng, P. (2022). Machine-learning-driven digital twin for lifecycle management of complex equipment. IEEE Transactions on Emerging Topics in Computing, 10(1), 9-22.

[13] Rolf, B., Jackson, I., Müller, M., Lang, S., Reggelin, T., & Ivanov, D. (2023). A review on reinforcement learning algorithms and applications in supply chain management. International Journal of Production Research, 61(20), 7151-7179.

[14] Sun, W., Xu, N., Wang, L., Zhang, H., & Zhang, Y. (2020). Dynamic digital twin and federated learning with incentives for air-ground networks. IEEE Transactions on Network Science and Engineering, 9(1), 321-333.

[15] Wang, J., Li, X., Wang, P., & Liu, Q. (2024). Bibliometric analysis of digital twin literature: A review of influencing factors and conceptual structure. Technology Analysis & Strategic Management, 36(1), 166-180.

[16] Xu, H., Wu, J., Li, J., & Lin, X. (2021). Deep-reinforcement-learning-based cybertwin architecture for 6G IIoT: An integrated design of control, communication, and computing. IEEE Internet of Things Journal, 8(22), 16337-16348.

[17] Dai, Y., Zhang, K., Maharjan, S., & Zhang, Y. (2020). Deep reinforcement learning for stochastic computation offloading in digital twin networks. IEEE Transactions on Industrial Informatics, 17(7), 4968-4977.

[18] Sun, W., Lei, S., Wang, L., Liu, Z., & Zhang, Y. (2020). Adaptive federated learning and digital twin for industrial Internet of things. IEEE Transactions on Industrial Informatics, 17(8), 5605-5614.

[19] Ivanov, D., & Dolgui, A. (2021). A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. Production Planning & Control, 32(9), 775-788.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Apr 19, 2024
Accepted: May 27, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 250–258, DOI 10.12694/scpe.v26i1.3786

A STUDY ON THE EFFECT OF DEEP REINFORCEMENT LEARNING IN CULTIVATING ATHLETE DECISION BEHAVIOR AND PSYCHOLOGICAL RESILIENCE

SHUYING SONG[*]AND KUN QIAN[†]

Abstract. In order to explore the relationship between the psychological resilience level and risk decision-making behavior of volleyball players, the author proposes a study on the effect of deep reinforcement learning in the cultivation of athlete decision-making behavior and psychological resilience. A survey and analysis were conducted on the psychological resilience level and risk decision-making behavior of 64 volleyball club athletes (29 males and 35 females) using the Psychological Resilience Inventory (PPI-A) and Sports Scenario Risk Decision Questionnaire. Construct a random forest regression model based on questionnaire data. The results indicate that there is a significant difference in risk decision-making behavior between athletes with high and low levels of psychological resilience in terms of benefits and losses =4.700,P=0.017,=22.065,P=0.000; There is a significant difference in risk decision-making behavior between athletes with high and low levels of psychological resilience when risk preference loss occurs =4.351,P=0.024, and in the context of positive and negative framing effects, the level of psychological resilience has no significant impact on decision-making behavior. The risk decision-making behavior of volleyball players is influenced by the framing effect, with negative framing and preference loss resulting in more risky behavior and a preference reversal; The level of psychological resilience affects the risk decision-making behavior of athletes in stressful situations, and athletes with high levels of psychological resilience have more adventurous behaviors.

Key words: Volleyball player, Psychological resilience, Framework effect, Risk decision-making, preference reversal

1. Introduction.

In 2019, the General Office of the State Council officially issued the ”Outline for Building a Sports Strong Country”, deploying the promotion of the construction of a sports strong country and fully leveraging the important role of sports in the new journey of building a socialist modernized strong country, and proposed five strategic tasks [1]. The second aspect is to enhance the comprehensive strength of competitive sports and enhance the ability to bring glory to the country by establishing a modern competition system with Chinese characteristics and promoting the development of professional sports [2].

In modern competitive sports, athletes making quick and accurate decisions during competitions is one of the key to achieving victory. However, with the tension and pressure of the competition, the decision-making ability and psychological state of athletes may be affected, which in turn can affect their competitive performance. Therefore, how to cultivate the decision-making ability and psychological resilience of athletes has become a focus of attention for many coaches and researchers [3]. Deep reinforcement learning, as an important branch of artificial intelligence, has shown tremendous potential in various fields, including gaming, finance, and healthcare. Its method of learning optimal strategies through interaction with the environment based on intelligent agents provides a new approach to solving the optimization of athlete decision-making behavior. Meanwhile, psychological resilience, as an important psychological trait possessed by athletes when facing pressure and challenges, is inherently linked to the concept of deep reinforcement learning [4]. Whether athletes can have strong psychological qualities to withstand pressure at critical moments in competitive competitions, and be able to maximize their learned skills, reduce mistakes, and steadily demonstrate their rightful technical level is an important factor in winning. In the face of difficulties and obstacles, having excellent sports skills is a prerequisite, and strong psychological qualities can support the performance of athletes' skills. Moreover, any skill in competitive sports requires years and months of accumulation to be acquired. Therefore, the cultivation of athletes requires starting from the grassroots level to grasp the spirit of work style, perseverance, and never giving up [5].

[*] Jilin Institute Of Physical Education, ChangChun, 130022, China

[†] College of Sports Science, Shenyang Normal University, Shenyang, 110034, China (Corresponding author, qiankun20232@163.com)

2. Literature Review. Existing research has confirmed that students who participate in physical exercise for a long time have higher levels of psychological resilience. The impact of different exercise intensities on psychological resilience also varies, regular participation in moderate intensity exercises leads to better psychological resilience. Cai et al. proposed a cloud edge device computing offloading method based on multi-agent deep reinforcement learning (MADRL), aimed at meeting various requirements of different tasks [6]. Sacchi, N. et al. proposed an essential redundant robot fault diagnosis and control scheme based on deep reinforcement learning (DRL) method combined with a set of sliding mode observers [7]. Wang, B. et al. used deep reinforcement learning (DRL) methods to actively control the flow of elliptical cylinders. The results indicate that DRL can learn active control strategies for the current configuration [8]. Eryilmaz, A. et al. investigated variables related to psychological resilience in pre adolescent youth, which is a trainable skill associated with psychological and academic performance [9]. Psychological resilience needs to have characteristics such as traits, processes, differences, outcomes, and dynamics, and is not a single or independent concept, but a three-dimensional and multi-level concept.

The author employs a questionnaire survey methodology to delve into the psychological resilience levels and risk decision-making behaviors among elite volleyball athletes in China. The aim is to establish a deeper understanding of these aspects and to explore potential correlations between them. The ultimate goal is to furnish valuable insights that can inform and enhance sports training programs and competitive performance strategies for these athletes.

3. Research Methods.

3.1. Research Object. Two men's volleyball teams (A team ranked third and B team ranked ninth) and three women's volleyball teams (C team ranked first in the league, D team ranked fifth, and E team ranked eighth) participated in the Chinese Volleyball League, with a total of 64 athletes. Among them, there are 29 males and 35 females; 52 strong generals, 12 at the first level; 31 active and previous national team athletes, 33 local team athletes; The average age is 22 years old; The average training period is 9 years. All athletes voluntarily participate in this survey study.

3.2. Research Methods. The author conducted a survey on 65 male and female athletes from 5 volleyball teams using the Psychological Resilience Inventory (PPI-A) and the Sports Scenario Risk Decision Questionnaire. The questionnaire is distributed and filled out by the relevant scientific researchers of the sports team. Before filling it out, the responsible scientific researchers explain in detail to the athletes the precautions for filling out the questionnaire. The questionnaire will be distributed one month after the end of the 2013-2014 season and will be collected one week later. A total of 65 questionnaires were distributed in this survey, with a response rate of 100%. Among them, 1 invalid questionnaire and 64 valid questionnaires were distributed. Finally, 64 athletes were included in this study. Then, according to the scoring criteria of the Psychological Resilience Scale, the psychological resilience score of each athlete is calculated, and the top 25% and bottom 25% are selected as the high and low psychological resilience groups in descending order of questionnaire scores. Next, statistical analysis was conducted on the framework effect and risk preference scores of athletes in the high and low psychological resilience groups in the four questions of the sports scenario risk decision-making questionnaire, in order to explore the relationship between psychological resilience level and risk decision-making behavior [10].

3.2.1. Research tools.

(1) Psychological Resilience Inventory (PPI-A). This questionnaire was developed by Colby et al. based on the Physical Performance Inventory (PPI) and is currently one of the main tools used by foreign researchers to evaluate exercise psychological resilience. It includes four dimensions: determination, self-confidence, positive cognition, and visual representation, with internal consistency coefficients of 0.86, 0.84, 0.82, and 0.75, respectively. The scale consists of 14 items, and the sum of the scores from the 4 dimensions is the total score of psychological resilience. This study translated the questionnaire into English and Chinese English by professional English teachers, and consulted two sports psychology experts to determine the final translation version of the questionnaire. Then, 20 college students were selected as the survey subjects, and two questionnaire evaluations were conducted with a one month interval before and after, and their retest reliability was tested. The results showed a retest reliability of 0.74 for both measurements.

Table 4.1: Four dimensional scores of psychological resilience for different teams (M±SD)

	man (n=29)		woman (n=35)		
	Team A (n=20)	Team B (n=9)	Team C (n=10)	Team D (n=15)	Team E (n=10)
determination	11.52 ± 2.11	11.35 ± 2.07	11.83 ± 1.20	10.21 ± 2.34	10.36 ± 1.20
Bao Confidence	15.16 ± 2.37	14.12 ± 1.63	15.48 ± 1.17	14.00 ± 2.07	12.56 ± 2.06
positive perception	15.40 ± 2.71	13.04 ± 2.53	14.54 ± 0.75	13.21 ± 2.05	12.30 ± 1.24
visual imagery	10.10 ± 2.30	10.58 ± 1.78	10.36 ± 1.73	10.57 ± 2.04	11.16 ± 2.01

(2) Sports Scenario Risk Decision Questionnaire. This questionnaire has a total of 12 test questions, each with only two choices: A and B, using the forced choice method of either choice, among them, there are 3 test questions about reference point effect, 4 test questions about framework effect and risk preference, 1 test question about decreasing sensitivity, and 4 test questions about psychological account. The scenario design includes technical and tactical choices, competition selection, reward and punishment preferences, etc. Based on the research content and survey subjects, the author selected four questions: framework effects and risk preferences, and replaced the involved sports scenarios with problem scenarios in volleyball matches, while using the "insider" problem scenario [11].

3.2.2. Data Analysis. Use SPSS 17.0 statistical software to conduct statistical analysis on the questionnaire results, including conducting independent sample t-tests and one-way ANOVA on the psychological resilience levels between different sports teams [12]. Conduct a multiple factor analysis of variance on the influencing factors of psychological resilience level (gender, sports level, and whether the national team is an athlete); The chi square test was conducted on the risk decision-making behavior of athletes, as well as the relationship between psychological resilience and risk decision-making behavior, with a significance level of 0.05.

4. Result analysis.

4.1. Sports psychological resilience.

4.1.1. Psychological resilience of different teams. The independent sample t-test results showed that there was no significant difference in determination, self-confidence, and visual representation between team A and team B. Positive cognition (t=1.801, p=0.048) showed marginal significance, and team A's positive cognition was to some extent superior to team B. The results of the analysis of variance showed significant differences in the determination of psychological resilience (F=4.098, P=0.022), self-confidence (F=9.282, P=0.000), positive cognition (F=8.013, P=0.001), and total score (F=3.248, P=0.028). The results of multiple comparisons (LSD) showed that the determination, confidence, and positive cognition of Team C were significantly better than those of Team D and Team E. Team D had significantly better confidence than Team E. However, there was no significant difference in the visual representation between Team C, Team D, and Team E (Table 4.1).

4.1.2. Psychological resilience of athletes of different genders, sports levels, and national teams. The results of multivariate analysis of variance showed that gender had a significant main effect on positive cognition (F=12.057, P=0.001), with male athletes having significantly better positive cognition than female athletes; Whether there are significant differences in determination (F=6.687, P=0.011), self-confidence (F=4.025, P=0.036), positive cognition (F=6.542, P=0.011), and visual representation (F=6.525, P=0.012) among national team athletes; The main effect of exercise level is not significant; The interaction between gender and sports level is significant in the positive cognition (F=4.682, P=0.021) dimension. The interaction between national team athletes and sports level is significant in the determination (F=7.201, P=0.007), self-confidence (F=4.854, P=0.017), and visual representation (F=15.434, P=0.000) dimensions. The interaction between gender and national team athletes in the positive cognition (F=7.057, P=0.009) dimension is significant (Table 4.2).

Table 4.2: Comparison of psychological resilience among athletes of different genders, sports levels, and whether they are national team members (M±SD)

	man (n=29)	woman (n=35)	master sports- -man (n=52)	Class A (n=12)	national team (n=31)	local team (n=33)
Determination	11.40± 2.04	10.73± 2.00	11.21± 1.71	10.30± 2.80	11.32± 1.82	10.80± 2.04
Bao Confidence	14.84± 2.20	14.01± 2.08	14.47± 2.10	14.00± 2.47	14.52± 2.32	14.22± 2.00
positive perception	15.00± 2.73	13.36±1.68	14.24± 2.27	13.05± 2.17	14.52± 2.53	13.54± 2.60
visual imagery	10.24± 2.14	10.66±2.01	10.54± 2.01	10.18± 2.33	10.28± 2.31	10.67± 1.80

Table 4.3: Proportion of Risk Decision Behaviors of Athletes from Different Teams [n (%)]

	Team A(n=20)		Team B(n=9)		Team C(n=10)		Team D(n=15)		Team E(n=10)	
	adven- ture	conser- vative	adven- ture	conser- vative	adven- ture	conser- vative	adven- ture	conser- vative	adven- ture	conser- vative
Front frame	10(50)	10(50)	3(22)	6(56)	3(20)	7(60)	5(22)	10(56)	5(40)	5(40)
Negative framework	13(54)	7(24)	6(56)	3(22)	5(40)	5(40)	13(76)	2(02)	9(80)	1(10)
Preferential benefits	4(20)	16(70)	3(22)	6(56)	3(20)	7(60)	5(22)	10(56)	4(30)	6(50)
Preference loss	15(64)	5(14)	8(78)	1(11)	9(80)	1(10)	15(100)	0(0)	8(70)	2(10)

Note: Adventure=seeking risk; Conservative=risk avoidance. The following table is the same.

Table 4.4: Comparison of Risk Decision Behaviors of Athletes under Positive and Negative Framework Effects Scenarios [n (%)]

	adventure	conservative	amount	χ^2	P
Front frame	25(30)	37(48)	62		
Negative framework	45(61)	17(17)	62		
amount	70(45)	54(33)	124	12.587	0.000

Table 4.5: Comparison of Risk Decision Behaviors of Athletes under Risk Preference Benefit and Loss Scenarios [n (%)]

	adventure	conservative	amount	χ^2	P
Preferential benefits	18(20)	44(60)	62		
Preference loss	54(75)	8(03)	62		
amount	72(47)	52(31)	124	41.403	0.000

4.2. Risk decision-making behavior of athletes.

4.2.1. Risk decision-making behavior of athletes from five teams. Frequency statistics were conducted on the risk decision-making behaviors of five teams, and the results showed that in terms of positive and negative framing effects, athletes tended to adopt conservative strategies (50% to 70%) in positive situations, while in negative situations, athletes tended to adopt more adventurous strategies (50% to 90%); In terms of risk preference, athletes tend to adopt conservative strategies (60% to 80%) when benefiting, and more inclined towards adventurous strategies (75% to 100%) when losing (Table 4.3). The test results indicate that compared to positive framing effects and risk preference benefits, athletes exhibit more risky behavior when negative framing and risk preference losses occur, and the difference is significant, χ^2=12.587, P=0.000, χ^2=41.403, P=0.000 (Tables 4.4 and 4.5).

4.2.2. The impact of gender, sports level, and national team athlete status on risk decision-making behavior. The chi square test results indicate that there is no significant difference in risk decision-

Table 4.6: Risk decision-making behavior of national and local team athletes in the context of positive and negative framework effects [n (%)]

	Front frame		χ^2	P	Negative framework		χ^2	P
	adventure	conservative			adventure	conservative		
national team (n=31)	7(12)	24(66)			22(60)	9(18)		
local team (n=33)	18(44)	15(34)	6.750	0.022	23(60)	10(20)	0.012	0.801

Table 4.7: Impact of risk preference on risk decision-making behavior of athletes with different levels of psychological resilience [n (%)]

	Preferential benefits		Preference loss		χ^2	P
	adventure	conservative	adventure	conservative		
High psychological resilience	4(16)	11(62)	11(62)	4(27)	4.700	0.017
Low psychological resilience	2(02)	13(76)	15(100)	0(0)	22.065	0.000

Table 4.8: The impact of different levels of psychological resilience on risk decision-making behavior in preference for gains and losses [n (%)]

	Preferential benefits		Preference loss		χ^2	P
	adventure	conservative	adventure	conservative		
Preferential benefits	4(16)	11(62)	2(02)	13(76)	0.207	0.538
Preference loss	11(623)	4(16)	15(100)	0(0)	4.351	0.024

Table 4.9: Impact of Positive and Negative Information Frameworks on Risk Decision Behavior of Athletes with Different Levels of Psychological Resilience [n (%)]

	Front frame		Negative framework		χ^2	P
	adventure	conservative	adventure	conservative		
High psychological resilience	7(36)	8(42)	10(56)	5(22)	0.432	0.350
Low psychological resilience	5(22)	10(56)	8(42)	7(36)	0.432	0.350

making behavior between gender and sports level in positive and negative frame effects and risk preference benefit losses. Is there a significant difference in decision-making behavior among national team athletes in positive frame effects, χ^1=6.750, P=0.022 (Table 4.6, only showing significant results) [13,14].

4.3. The relationship between different levels of psychological resilience and risk decision-making behavior . The chi square test and random forest results indicate that there is a significant difference in risk decision-making behavior between athletes with high and low levels of psychological resilience in terms of benefits and lossesχ^2=4.700,P=0.017,χ^2=22.065, P=0.000 (Table 4.7); There is a significant difference in risk decision-making behavior between athletes with high and low levels of psychological resilience when risk preference loss occurs χ^2=4.351, P=0.024 (Table 4.8), but in the positive and negative framing effect scenario, the level of psychological resilience has no significant impact on decision-making behavior (Tables 4.9 and 4.10).

4.4. Random Forest Regression. The library function for random forest regression in MATLAB is also TreeBagger.

4.4.1. Original dataset. The original dataset contains 1017 samples and 37 feature variables. The minimum sample size L of the leaf node is used as the hyperparameter for the random forest regression model. When selecting hyperparameters for the model, the training set is used to train the random forest regression model, and the out of bag error value (MSE) returned by the model is used as the evaluation indicator, when

Table 4.10: The impact of different levels of psychological resilience on risk decision-making behavior under positive and negative information frameworks [n (%)]

	High psychological resilience		Low psychological resilience		χ^2	P
	adventure	conservative	adventure	conservative		
Front frame	7(36)	8(42)	5(22)	10(56)	0.128	0.708
Negative framework	10(56)	5(22)	8(42)	7(36)	0.128	0.708

MSE is at its minimum, L=14. After determining the hyperparameters, a random forest regression model was trained using the training set, and the performance of the model was tested using the test set. The results showed that MAE=0.1074,MSE=0.0164 R^2=0.6105, R^2_{adj}=0.3824. Using the entire dataset containing all 1017 samples, continue training the model to obtain the importance coefficients of feature variables and the random forest regression model [15]. The importance coefficients of 37 characteristic variables, the larger the value, the higher the importance. A negative value indicates that the importance is lower than 0. In this model, the top ten feature variables with the highest importance, from high to low, are positive coping, depression level, mindfulness level, health status, sleep quality, age, anxiety level, and stress level.

4.4.2. Filtering Datasets. The selected dataset contains 1017 samples and 8 feature variables. The out of bag error, MSE, is used for hyperparameter selection in the random forest regression model, when MSE is at its minimum, L=13. After determining the hyperparameters, the random forest regression model was trained using 916 samples from the training set, and the performance of the model was evaluated using 101 samples from the testing set. The results showed that MAE=0.1291,MSE=0.0239, R^2=0.6365, R^2_{adj}=0.6049. Finally, the entire dataset containing all 1017 samples was used to continue training the model, resulting in importance coefficients and a random forest regression model containing 8 feature variables. The importance coefficients and importance rankings of the 8 feature variables were ranked from high to low, followed by positive coping, depression level, mindfulness level, sleep quality, health status, and negative coping [16].

4.5. Psychological resilience levels of different teams. The results of this study show that the psychological resilience level of Team A is better than that of Team B, but there is no significant difference. The psychological resilience levels of Team C and Team D are both better than those of Team E. According to the results of the five teams in the past two years, Team A is in the top four of the league, Team B is in a middle and lower position, Team C is a top three team and won last year's league championship, and Team E is on the brink of relegation almost every season. Psychological resilience is an important factor affecting sports performance, and a higher level of psychological resilience is more conducive to improving sports performance.

4.6. The impact of gender, sports level, and national team athlete status on psychological resilience. The results of this study show that male athletes have significantly better positive cognition than female athletes, while general athletes have better psychological resilience than first level athletes, but there is no significant difference. National team athletes have better determination, self-confidence, and positive cognition than local team athletes, but there is no significant difference. Previous studies have suggested that psychological resilience is a dynamic development process, and excellent athletes have a higher level of psychological resilience. Generals and national team athletes have richer training and competition experiences, which continuously improve their psychological resilience, making them more confident, resilient, and resilient, and able to actively cope with training and competition problems, demonstrating obvious psychological advantages.

4.7. Risk decision-making behavior of athletes.

4.7.1. Overall risk decision-making behavior. People exhibit significant framing effects in risk scenarios and exhibit preference reversal in their behavioral choices. Previous studies have found that the risk decision-making behavior of team project athletes is significantly influenced by the framing effect. The results of this study also show that volleyball players exhibit risk avoidance in positive information and benefit situations, and risk seeking in negative and loss situations [17]. Moreover, compared to the positive framework and preference gain scenario, more athletes exhibit risky behavior and exhibit a significant preference reversal when the negative framework and preference loss occur. This indicates that the risk decision-making behavior of athletes is also influenced by the framing effect and exhibits preference reversal. In addition, athletes did not show a significant risk aversion tendency in the positive framework, but showed a significant risk seeking tendency in the negative framework. Previous studies have also found that athletes do not exhibit a risk preference in their decision-making behavior under a positive framework, but only exhibit a clear risk preference phenomenon under a negative framework. There are also studies that have found that college students tend to seek risk within a positive and negative framework, and there is no gender difference. In addition, this study found that athletes showed significant risk avoidance when receiving benefits and obvious risk seeking when receiving losses. Previous studies have found that college students showed significant risk avoidance when receiving positive returns, and there was no significant difference between risk avoidance and risk preference when receiving negative returns. Compared to the general population, athletes are a special group, and strict sports training shapes their unique personality traits. Especially for athletes, the unique social and cultural background has a significant impact on success and failure. Athletes are sensitive to positive information during training and competition, and are more sensitive to negative information. For example, success (reward) and failure are goals that athletes deliberately pursue and avoid, especially in order to avoid potential loss of benefits caused by failure, which in turn affects their decision-making behavior in competition [18]. In addition, athletes often face leading or falling behind situations, such as serving at key moments in volleyball matches. High quality serving can disrupt the opponent's pass and take the initiative, but it also increases the risk of serving errors. In this situation, different athletes may adopt different risk decision-making behaviors. Therefore, diverse scenarios (or gains and losses) provide athletes with more risk decision-making opportunities than ordinary people.

4.7.2. The impact of gender, sports level, and national team athlete status on risk decision-making behavior. The results of this study show that in a positive framework, female athletes exhibit higher levels of conservative behavior than male athletes, and lower levels of adventurous behavior than male athletes. In a negative framework, female athletes exhibit higher levels of adventurous behavior than male athletes, and lower levels of conservative behavior than male athletes; Women's risk-taking behavior is higher than that of male athletes in both negative frames and losses. For example, 74% of female athletes choose risk-taking behavior in negative frames, and 91% of female athletes choose risk-taking behavior when facing losses. Both male and female athletes are affected by the framing effect, but there is no significant difference. Previous studies have also found that both men and women tend to seek risk in both positive and negative frameworks, but there is no significant difference.

Women's risk seeking in negative frameworks is more pronounced and significantly different compared to positive frameworks. This may be due to women's higher sensitivity to language frameworks. But some studies have also found that female referees tend to be more adventurous in a positive framework, while men tend to be more conservative, and there is no difference between men and women in a negative framework. This may be due to women being more susceptible to stress events than men, and experiencing more conflicts, setbacks, and accompanying negative emotions, which can affect their decision-making behavior. There are also studies that have found that male athletes are significantly affected by the framing effect in the player problem, while female athletes are not affected by the framing effect. This may be because women have a natural tendency towards conservatism and lower risk-taking, thus leaning towards conservative behavior. From the above results, it can be seen that future research needs to further explore the gender effects of risk decision-making behavior. For the relationship between skill level and framing effect, whether it is in player or outsider problems, average level athletes are affected by framing effect, while excellent athletes are only affected by framing effect in player problems, and low-level athletes are affected by framing effect in positive framing situations. The results of this study also show that both first level athletes and top athletes are affected by the framing effect, but lower level first level athletes are slightly more affected by the framing effect than top athletes. Low level athletes are more susceptible to framing effects, which may be closely related to their sports experience [19].

General athletes have relatively less sports experience and less competition training, and their confidence differs significantly from that of excellent athletes. These factors affect their decision-making behavior, which may also be the reason why low-level athletes exhibit significantly higher risk-taking behavior in a positive framework than national team athletes. But some studies have also found that referee level does not affect the risk decision-making behavior of gymnastics referees.

In addition, this study also examined the decision-making behavior of national team athletes, and the results showed that national team athletes had more conservative decision-making behavior in positive and beneficial situations, especially in positive situations where conservative behavior was significantly higher than that of local team athletes. This may be related to the identity characteristics of national team athletes, who are more susceptible to the influence of success and benefits, less willing to fail, and therefore more inclined to avoid negative information. In a positive framework, they exhibit more conservative behavior.

4.8. Relationship between psychological resilience level and risk decision-making behavior . Regarding the relationship between psychological resilience level and risk decision-making behavior, research has found that in the context of positive and negative framing effects, compared with college students in the low psychological resilience group, college students in the high psychological resilience group are significantly more inclined to seek risk. This may be because individuals with high psychological resilience have better adaptability when dealing with stress in risk situations, can respond to risks and setbacks with a positive attitude, and are more inclined to seek risk. On the other hand, individuals with low psychological resilience tend to consider the potential losses of risk when facing risk situations, and therefore make higher assessments of risk and adopt risk avoidance. The results of this study showed that in the context of risk preference, athletes with high and low levels of psychological resilience had consistent decision-making tendencies when it came to risk preference. There was no significant difference in conservative behavior when it came to benefits, but there was a significant difference in risky behavior when it came to losses. Athletes with high levels of psychological resilience had more risky behaviors, namely risk seeking tendencies. This result is consistent with previous research, indicating that athletes with high levels of psychological resilience exhibit a clear risk seeking tendency in risk scenarios. This performance may be related to individuals with high levels of psychological resilience having better adaptability and recovery abilities when dealing with stress or setbacks. At the same time, athletes with high levels of psychological resilience can still demonstrate their technical and tactical skills even in difficult situations, maintaining sustained self-confidence, focus, and control over stress in stressful situations. These may also be important reasons why high resilience athletes exhibit more risky behavior in negative risk decision-making situations.

In addition, emotions influence people's risk decision-making behavior in stressful situations. Research has found that individuals with higher levels of positive emotions accumulate more resources such as psychological resilience, making them more likely to make optimistic judgments and exhibit higher risk seeking preferences, while negative emotions such as anxiety levels show a significant negative correlation with risk decision-making, making them more likely to make pessimistic judgments and exhibit higher risk avoidance preferences. This is because anxiety promotes pessimistic evaluations of future events [20].

Individuals with high levels of psychological resilience often have more positive emotional experiences because they make athletes more outgoing and communicative, allowing them to remain relaxed and calm, and to be competitive in many situations. They also have lower levels of anxiety, a better perception of self-confidence, and an unshakable belief in their ability to control adversity, making them more conducive to making good decisions in stressful situations. Due to the fact that positive mentality is one of the important characteristics of sports psychological resilience, athletes with high levels of psychological resilience have lower levels of negative emotions, which may be an important reason for differences in risk decision-making behavior among athletes with different levels of psychological resilience.

5. Conclusions.

1. The overall psychological resilience level of the top ranked teams is better than that of the bottom ranked teams;
2. The risk decision-making behavior of volleyball players also follows the framework effect of prospect theory, and when negative frameworks and risk preference losses occur, athletes exhibit higher risk taking behavior, leading to a phenomenon of preference reversal;
3. The influence of gender and sports level on the risk decision-making behavior of volleyball players is not significant, and national team athletes have more risk seeking tendencies;
4. Psychological resilience affects the risk decision-making behavior of volleyball players, and volleyball players with high levels of psychological resilience tend to engage in more risky behaviors, namely risk

seeking tendencies.

REFERENCES

[1] Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. Neuron, 107(4), 603-616.

[2] Tsai, W. L., Su, L. W., Ko, T. Y., Pan, T. Y., & Hu, M. C. (2021). Feasibility study on using ai and vr for decision-making training of basketball players. IEEE Transactions on Learning Technologies, 14(6), 754-762.

[3] Puram, P., Roy, S., Srivastav, D., & Gurumurthy, A. (2023). Understanding the effect of contextual factors and decision making on team performance in Twenty20 cricket: an interpretable machine learning approach. Annals of Operations Research, 325(1), 261-288.

[4] Rossi, A., Pappalardo, L., & Cintia, P. (2021). A narrative review for a machine learning application in sports: an example based on injury forecasting in soccer. Sports, 10(1), 5.

[5] Li, B., **e, K., Huang, X., Wu, Y., & **e, S. (2022). Deep reinforcement learning based incentive mechanism design for platoon autonomous driving with social effect. IEEE Transactions on Vehicular Technology, 71(7), 7719-7729.

[6] Cai, J. , Fu, H. , & Liu, Y. . (2023). Multitask multiobjective deep reinforcement learning-based computation offloading method for industrial internet of things. IEEE internet of things journal, 9(1), 206-227.

[7] Sacchi, N. , Incremona, G. P. , & Ferrara, A. . (2023). Sliding mode based fault diagnosis with deep reinforcement learning add-ons for intrinsically redundant manipulators. International Journal of Robust and Nonlinear Control(15), 33.

[8] Wang, B. , Wang, Q. , Zhou, Q. , & Liu, Y. . (2022). Active control of flow past an elliptic cylinder using an artificial neural network trained by deep reinforcement learning. Applied Mathematics and Mechanics, 43(12), 1921-1934.

[9] Eryilmaz, A. , Hacer Yldrm-Kurtulu, & Doenyas, C. . (2023). Positive affect, negative affect, and psychological resilience mediate the effect of self-compassion on mental toughness: a serial mediation analysis. Psychology in the schools, 33(6), 627-643.

[10] Yasar, O. M., & Turgut, M. (2020). Mental toughness of elite judo athletes. Acta Med, 36(2), 995-8.

[11] Ajilchi, B., Mohebi, M., Zarei, S., & Kisely, S. (2022). Effect of a mindfulness programme training on mental toughness and psychological well-being of female athletes. Australasian Psychiatry, 30(3), 352-356.

[12] Cooper, K. B., Wilson, M. R., & Jones, M. I. (2021). Fast talkers? Investigating the influence of self-talk on mental toughness and finish times in 800-meter runners. Journal of Applied Sport Psychology, 33(5), 491-509.

[13] Graham, S. M., Martindale, R. J., McKinley, M., Connaboy, C., Andronikos, G., & Susmarski, A. (2021). The examination of mental toughness, sleep, mood and injury rates in an Arctic ultra-marathon. European journal of sport science, 21(1), 100-106.

[14] Hunt, M. Q., Novak, C. E., Madrigal, L. A., & Vargas, T. M. (2020). Strategies for develo** mental toughness in high school athletes. Strategies, 33(1), 14-19.

[15] Murray, R. M., Dugdale, J. H., Habeeb, C. M., & Arthur, C. A. (2021). Transformational parenting and coaching on mental toughness and physical performance in adolescent soccer players: The moderating effect of athlete age. European Journal of Sport Science, 21(4), 580-589.

[16] Eraña-Díaz, M. L., Cruz-Chávez, M. A., Rivera-López, R., Martínez-Bahena, B., & Cruz-Rosales, M. H. (2020). Optimization for risk decision-making through simulated annealing. IEEE Access, 8, 117063-117079.

[17] Markiewicz, Ł., Muda, R., Kubińska, E., & Augustynowicz, P. (2020). An explanatory analysis of perceived risk decision weights (perceived-risk attitudes) and perceived benefit decision weights (perceived-benefit attitudes) in risk-value models. Journal of Risk Research, 23(6), 739-761.

[18] Søbjerg, L. M., Taylor, B. J., Przeperski, J., Horvat, S., Nouman, H., & Harvey, D. (2021). Using risk factor statistics in decision-making: prospects and challenges. European Journal of Social Work, 24(5), 788-801.

[19] Guan, H., Dong, L., & Zhao, A. (2022). Ethical risk factors and mechanisms in artificial intelligence decision making. Behavioral Sciences, 12(9), 343.

[20] de Leeuw, A. W., van der Zwaard, S., van Baar, R., & Knobbe, A. (2022). Personalized machine learning approach to injury monitoring in elite volleyball players. European journal of sport science, 22(4), 511-520.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Apr 23, 2024
Accepted: Jun 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 259–268, DOI 10.12694/scpe.v26i1.3815

THE PERSONALIZED LEARNING PATHS FOR DIGITAL MEDIA TECHNOLOGY EDUCATION BASED ON BIG DATA

PENG CHANGRONG[*], LI QI[†], ZHANG XIAODONG[‡], AND SHA HAIYAN[§]

Abstract. The paper intends to study the evolution of domain knowledge by studying the spatial-temporal collaborative model. A joint knowledge network model based on the time-space domain is proposed to represent the knowledge base. The skeletal clustering algorithm analyzes the evolution of knowledge networks over the years. According to the concept of the evolution process of knowledge, the paper makes a connection and path analysis of its evolution track. An empirical study of the digital media field is carried out. The results show that the algorithm proposed in this paper can extract the evolution trajectory of domain knowledge that varies with year. The path of knowledge evolution can show the correlation between research topics, hot topics, core literature, the evolution law of research topics and research methods of multiple disciplines, and the cross-characteristics of multiple disciplines.

Key words: Knowledge evolution; Evolutionary pathway; Spatio-temporal correlation; Learning path; Digital media technology

1. Introduction. In the big data environment, various disciplines are developing rapidly, and research papers in various disciplines are increasing rapidly. Therefore, an accurate and practical grasp of the trajectory of knowledge generation, development, evolution, and extinction is helpful for researchers to grasp the research focus of this field accurately and quickly find its key and frontier problems. Identifying the research hotspot and its evolution law in this field can realize efficient resource allocation of scientific research, support scientific decision-making and promote scientific innovation. In this context, it is essential to determine the path of knowledge evolution in the subject area.

At present, many scholars have used different methods to study the generation and visualization effect of the knowledge evolution path. The primary path method based on the citation network measures the connectivity of a loop less network by using the neutrality between nodes composed of the nodes with the most significant number of vertices. The primary research method is to use the global connectivity of the citation network to extract the primary way and to study the development trend and core literature, essential people and significant events in the academic field. The primary path analysis method is used to explore the data enveloping analysis method.

Literature [1] explores the knowledge diffusion path in data quality based on primary path analysis. Literature [2] uses main path and edge clustering methods to identify the dominant knowledge flow and activity orientation in science and technology adoption research. Literature [3] uses the method of three main channels to initially show the main clues of knowledge evolution in mobile libraries. Literature [4] shows its rich disciplinary evolution characteristics by extracting the agglutinite and structural cave groups associated with the primary pathways. Literature [5] proposes a research idea based on title clusters, that is, by clustering and classifying keywords to extract hidden information in the text. Many researchers have adopted the method of coward-based clustering to study knowledge evolution. Literature [6] uses time-series mapping technology to express and analyze topics dynamically based on clustering. Literature [7] uses the LDA method to represent topics' deep semantic characteristics and constructs the topics' evolution trajectorypics in each period. Literature [8] studies the evolution trajectory of topics and proposes a symbiotic network of topic probability distribution based on weights. The co-word clustering method based on SciMAT was used to carry out the

[*] College of Art, Hebei University of Economics and Business, Shijiazhuang, 050061, China
[†] College of Arts, Cheongju University, Cheongju, 28503, Korea (Corresponding author, liqi20231113@163.com)
[‡] College of Art, Hebei University of Economics and Business, Shijiazhuang, 050061, China
[§] College of Art, Hebei University of Economics and Business, Shijiazhuang, 050061, China

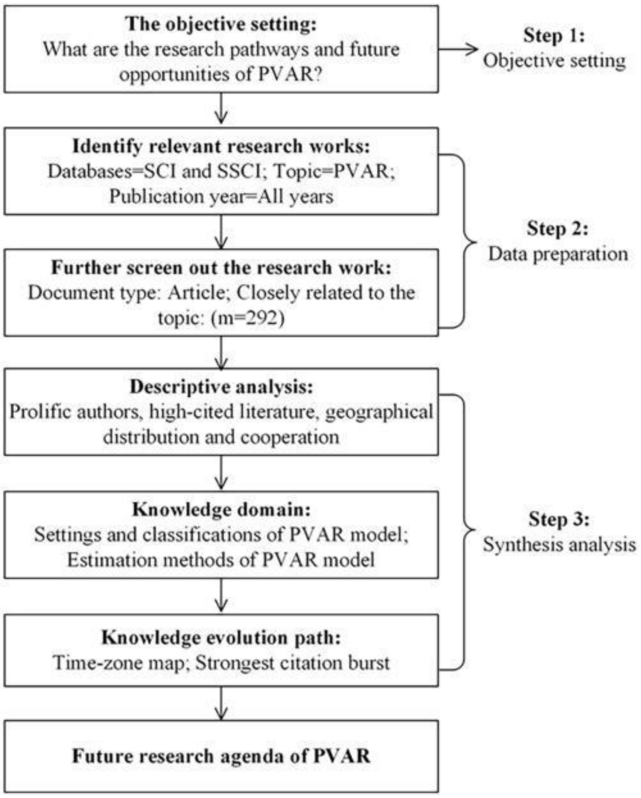

The objective setting:
What are the research pathways and future opportunities of PVAR?
Step 1:
Objective setting
Identify relevant research works:
Databases=SCI and SSCI; Topic=PVAR; Publication year=All years
Further screen out the research work:
Document type: Article; Closely related to the topic: (m=292)
Step 2:
Data preparation
Descriptive analysis:
Prolific authors, high-cited literature, geographical distribution and cooperation
Knowledge domain:
Settings and classifications of PVAR model; Estimation methods of PVAR model
Knowledge evolution path:
Time-zone map; Strongest citation burst
Step 3:
Synthesis analysis
Future research agenda of PVAR

Fig. 2.1: *Reflections on the way of knowledge evolution.*

trajectory diagram of the dynamic evolution of ISLS research hotspots in literature [9], the Jaccard coefficient measured the similarity of topics, and the current ISLS research hotspots were determined by constructing the evolution trajectory of topics. Literature [10] establishes a spatio-temporal scale-oriented domain knowledge evolution situation analysis model and connects it with the knowledge evolution process. The existing research ignores the non-key academic resources in the citation network, and it isn't easy to dig out the evolution law of the topic from the evolution track of academic resources.

This project intends to study the research idea of knowledge evolution in the time-space domain: the classical knowledge network is taken as the research object, and the shortest circuit with the best skeleton characteristics is taken as its evolutionary track to construct a complete knowledge evolution context in time-space scale [11]. This project intends to take academic papers in the digital media industry in CNKI as the primary research object by constructing an annual time series and extracting evolutionary trajectories based on bone clustering to study and analyze the development process.

2. Model framework. The research ideas of this paper can be summarized in Figure 2.1. This paper consists of four stages: data collection, primary path analysis, concept integration and knowledge evolution path construction. Firstly, the paper extracts "title," "keywords," and "abstract." The pauses in the sample were reduced and "dried." Then, the topic modeling method extracts keywords and builds a vocabulary dictionary. This paper uses isi.exe and CitNetw.exe to process data. The global mainstream route method (GMP) is selected, and the path between each junction and each node is extracted by an exhaustive method [12]. The path with the highest arc weight is the main line, and its literature is regarded as the essential reference material. The document-subject word matrix is established. The vector space modeling method is used to measure the similarity of the thesis. They are sorted according to their similarity with core papers to form appropriate core document nodes. In the framework of SATI, elements such as "keywords," "titles," and "abstracts" are extracted in a particular order, and then semantic mining of each node is carried out through topic modeling to realize the expression of each node topic in the text. Finally, the paper describes the research links in the text, thus establishing a text-based approach to knowledge evolution.

2.1. Knowledge network module. A knowledge network is essential for studying knowledge development in knowledge graphs. A knowledge network comprises nodes and edges, in which nodes represent knowledge elements and edges represent knowledge connections between entities. According to the difference of entity units, the nodes can be papers, patents, books, keywords, etc. According to the relevance of knowledge, the edge can be divided into reference relationships, symbiotic relationships, and cooperative relationships [13]. In this method, domain keywords are treated as nodes and evolutionary weights as edges, which are modified. Compared with reference networks, co-lexical networks can reflect the evolution of entity concepts in networks more directly and efficiently. The knowledge network established in this project is a weighted undirected network, and its research content is divided into two parts: The first is to evaluate this kind of knowledge network and its importance on the network diagram, such as keyword frequency, node degree, intermediate centrality, etc. The second is the shortest path, critical path and average path length are studied based on node connection. Network node analysis is often used to obtain the distribution of network topics, while network path analysis is used to predict the development direction of domain knowledge and find research hotspots.

2.2. Knowledge base vocabulary extraction. Research on Chinese vocabulary extraction has dramatically progressed, and relatively perfect methods have been developed. Among them, the Chinese automatic identification system NLPIR of the University of Chinese Academy of Sciences is the representative. This section discusses using the NLPIR to extract terms and data from files. Firstly, the text base related to a specific domain is collected, and the Key Extract Get Keywords algorithm in NLPIR is used to extract the vocabulary of a single text and store it in HashMap as a key-value pair. The key represents the keyword [14]. The value indicates the frequency of keyword occurrence and the number of keywords is collected. The first step is to extract the file, store it in the hash map when the first keyword appears, and set the key value to 1. If the keyword already appears in the hash map, the corresponding value of the keyword is increased by 1. Keywords in all files up to the current year are calculated. Finally, the keywords are sorted in descending order according to the value, and the N keywords with the highest frequency are listed as the domain words collection. The process of lexical extraction is shown in Figure 2.2. (The image is quoted in Intelligent RFQ Summarization Using Natural Language Processing, Text Mining, and Machine Learning Techniques).

2.3. Construction of time-space joint knowledge ne twork. The spatiotemporal collaborative knowledge network describes the change of knowledge with time in a dynamic process by establishing the spatiotemporal knowledge network. This project divides the collaborative knowledge network into two stages: generating a new knowledge network every year and building a continuous network based on multiple network nodes that have emerged in the last year [15]. The core work of knowledge network construction is to extract the relationship weights among network nodes. An evolutionary relationship can be regarded as an inter-entity relationship, depending on the text's lexical meaning and occurrence times. Over time, the degree of evolution has increased. This paper proposes the definition of evolutionary relation, that is, given file K, document knowledge concept entity sequence is represented as $R = \{r_1, r_2, r_3, \ldots\}$, and formula (2.1) is used to calculate the semantic distance between two entity concepts r_i and r_j in series R.

$$\operatorname{dis}(r_i, r_j) = \sum_{T_i \in R} |j - i|/n$$

i and j represent a particular class of knowledge in the series. And n represents a pair of information in a particular category in the series. The lower the semantic distance, the higher the degree of evolution between the two concepts r_i and r_j. A meaningful distance threshold σ is set in the test. If the relative distance between two information pairs is more significant than a particular critical value, they are evolutionarily related; otherwise, they are not evolutionarily related. If the node r_i is evolutionarily related to the node r_j, then there must be a connection path in the knowledge network of node pairs. The definition of evolutionary distance is given in

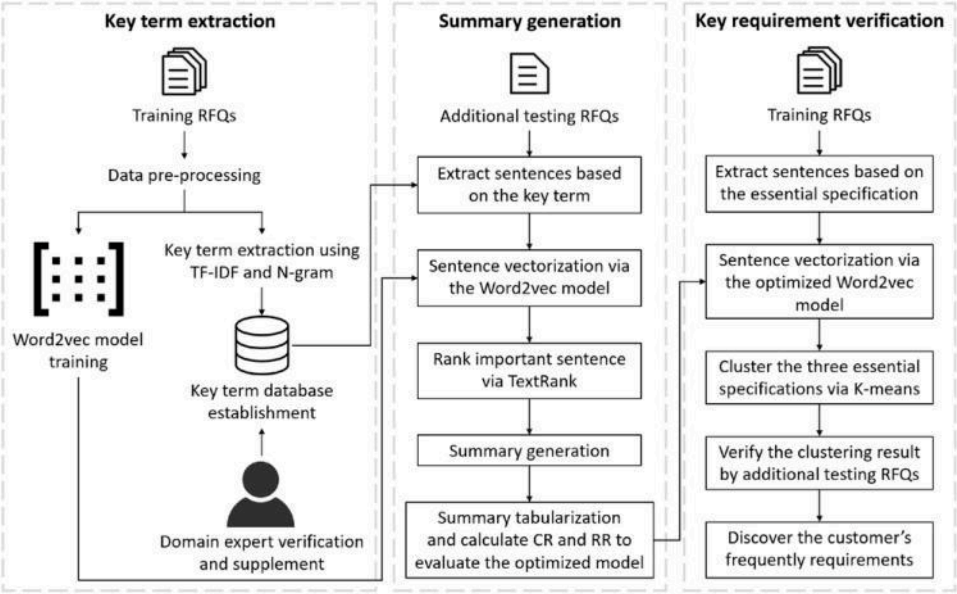

Key term extraction
Training RFQs
Data pre-processing
Word2vec model training
Key term extraction using TF-IDF and N-gram
Key term database establishment
Domain expert verification and supplement
Summary generation
Additional testing RFQs
Extract sentences based on the key term
Sentence vectorization via the Word2vec model
Rank important sentence via TextRank
Summary generation
Summary tabularization and calculate CR and RR to evaluate the optimized model
Key requirement verification
Training RFQs
Extract sentences based on the essential specification
Sentence vectorization via the optimized Word2vec model
Cluster the three essential specifications via K-means
Verify the clustering result by additional testing RFQs
Discover the customer's frequently requirements

Fig. 2.2: *Term extraction process.*

formula (2.2):

$$\operatorname{evo}\left(r_i r_j\right)=\exp \left(\left[\sum_{r_i r_j \in \bar{K}} \operatorname{dis}\left(r_i, r_j\right)\right)^2 /\left(2 m^2 \gamma^2\right)\right] / m^2$$

$\bar{K}$ represents the set of documents in which the entity concepts r_i and r_j exist. m represents the number of occurrences and $\operatorname{evo}(r_i r_j)$ has a low value, indicating that the evolution from r_i to r_j is relatively simple. The detailed process of extracting evolutionary relationships is shown in Figure 2.3. Then, the extracted keywords are input into the NLPIR automatic segmentation software to form a customized dictionary so that the software can be coarse granular segmentation [16]. Segmentation of individual text. User-defined words are filtered in the segmentation results to initialize keywords in the file. This is then combined with the contiguous neighboring keywords in the series to obtain the new nonoverlapping sequence R'. Then, the relationship between the keyword pairs in the sequence R' is analyzed. For example, there are two keywords r_i and r_j in R', both of which are stored in the form $\{r_{ij}, e_{ij}, n_{ij}\}$. Where r_{ij} is a pair of associations, e_{ij} is the semantic spacing of associations in a file, and n_{ij} is the number of occurrences of a pair of associations. The association pairs in the literature were statistically analyzed. n_{ij} and n_{ij} were accumulated for the association pairs that occurred frequently. Finally, the average semantic distance and occurrence time between the two groups of associations are obtained. Formula (2.2) calculates the evolutionary distance between the correlations and determines the weights between the correlations.

A knowledge network based on the time-space scale is established with keywords as nodes and evolutionary distance as weights. Figure 2.4 shows the joint knowledge network on a time-space scale over three years. The dots represent the knowledge of something. The larger the diameter of the circle center, the more critical the information held in the network [17]. The connections between nodes represent evolutionary connections. When the weight value is low, the nodes in the network are close to each other, indicating that the evolution of the two is higher. The dotted lines show the intersecting views of knowledge in each year's knowledge network. These repeated knowledge concepts establish the knowledge evolution relationship between successive years.

2.4. Skeleton cluster analysis. This paper focuses on extracting an optimal evolutionary route from this knowledge network. An ideal evolutionary trajectory can be regarded as the interconnection of multiple

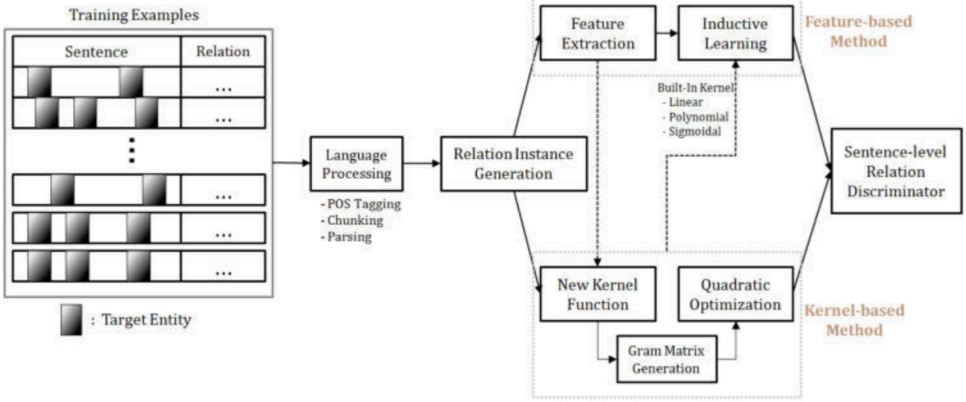


Fig. 2.3: *Evolutionary relationship extraction process.*

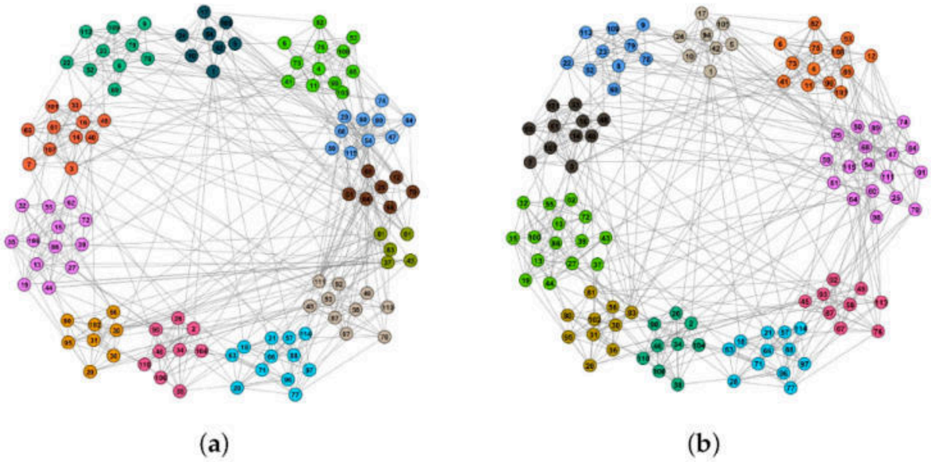


Fig. 2.4: *Space-time joint knowledge network structure.*

network systems, and the framework of the network system plays a role in supporting the network system, so the perfect network system must have the characteristics of centrality and connectivity. A knowledge network model based on "bone clustering" is established. Its general idea is: "partial aggregation, global correlation." Each cluster can be regarded as a knowledge topic, and its core node should be dispersed among multiple knowledge topics as much as possible. This core node can become the clustering center of a topic to achieve the best topic clustering. "Overall association" connects all nodes in each hierarchy to form an overall framework. In principle, the whole structure should be able to cover the entire knowledge network fully and simultaneously, ensuring that the sum of topic clustering effects of each part reaches the maximum.

The shortest path is seen as the best evolutionary path. The shortest path refers to the minimum cost required to get from one node to another, which can be regarded as the shortest path from one node to another [18]. Because the shortest circuit corresponds to the starting and ending points of evolution, it is necessary to use the clustering method of the molecular evolutionary tree to study the importance of various shortest circuits in the evolutionary process. A node cluster comprises multiple nodes; one node should be a good centrality, and the neighboring node is the cluster's core, thus forming a knowledge body. The detailed expression of the node clustering coefficient is expressed in the expression (2.3):

$$CH(r)=\sum_{z_i \in Z} djs\left(z_i, r\right)/Z_n$$

CH(r) is the clustering coefficient in trunk node r; Z is the subject related to r; Z_n is the number of nodes included in topic Z; Where dis$\left({}^*,{}^*\right)$ is the shortest between the nodes. When CH(r) cluster coefficient is the lowest, r is regarded as the core of the topic cluster. The whole skeleton was clustered and the best skeleton was selected. A specific calculation formula is expressed in equation (2.4):

$$SH(R)=\sum_{r_i \in R} CH\left(r_i\right)/Z_n$$

Z_n is the number of backbone nodes included in the backbone R. When the average cluster factor $SH(R)$ in a chain R is the lowest, the chain it represents is the optimal evolutionary chain.

3. Experimental research.

3.1. Experimental data. This paper selects the field of digital media as a new subject to conduct empirical research combined with the development of the subject and the current hot research direction. This research followed the following procedures: First, CNKI was selected, and the search keywords were ”media” and ”digital media.” This paper analyzes academic papers from 1997 to 2022 using ”abstract” and ”keyword” as search items. Articles in CAJ format are then downloaded by year and stored as the corresponding annual directory for ”1997-01”. If many papers appear in a given year, 300-500 papers are selected as the standard according to the number of paper downloads and citations. Secondly, using the ”save as” function provided by CAJ Viewer, the conversion of the CAJ file to a TXT file is realized to facilitate the running of the Java program. Some of the papers published in the early stages are stored in images, so the process will soon produce confusing code. It selects 5,214 typical papers published between 1997 and 2022. These articles can reflect the current development trend and research progress of digital media.

3.2. Test results. The experimental part is based on the knowledge network and integrates the academic papers published in digital media for many years to establish a complete subject knowledge network. According to the word frequency and node degree of the network, it is comprehensively sorted out. This paper analyzes the evolution of the knowledge of the digital media knowledge network over the years and extracts the evolution path to show the development course of the digital media field.

Firstly, NLPIR automatic segmentation technology is used to select 10 keywords for each text search result. After analyzing all the keywords in the text and their corresponding frequencies, 953 keywords with higher frequencies are selected and named ”digital media.” Table 3.1 lists the 10 most frequently used keywords, among which ”digital media,” ”media,” and ”traditional media” are all words with vital representative significance in digital media. In a sense, this illustrates the effect of these words’ extraction.

Integrating all digital media publications from 1997 to 2023 establishes and studies a complete knowledge network covering 27 years at the nodal level. A database of 953 knowledge words was formed, and the order of related topics was extracted from the literature [19]. Use formula (2.2) to treat keywords as network nodes. The evolutionary distance is the network edge that establishes the knowledge network.

Node degree reflects the number of associations of nodes in the knowledge network. The more associations there are, the higher the importance of the keyword. Figure 3.1 shows the 116,274 knowledge relationships corresponding to 953 keywords, in which the degree of keywords is distributed in a long tail, indicating that there is a small part of the core knowledge with high nodal, while most of them are low nodal, and this information are the ”Bridges” connecting this information.

Table 3.2 lists an index of some of the top 20 subject words of the year. Through the statistics of the information in the list, it can be found that in the news reports of 1997, the keywords are ”TV,” ”radio,” and ”audiovisual teaching.” After 2007, ”network,” ”Internet,” ”mobile phone,” and other keywords have appeared in people’s vision. Keywords such as ”media” and ”TV” are widely representative in the field and frequently appear yearly. It also reflects the evolving digital media landscape with the times.

Table 3.1: *List of top ten keywords in integrated word frequency in digital media.*

Serial number	Keyword	Occurrence frequency
1	Digital media	871
2	media	738
3	Old media	695
4	Intel	565
5	news	426
6	diffuse	371
7	Digital television	350
8	network	339
9	advertisement	326
10	TV	310

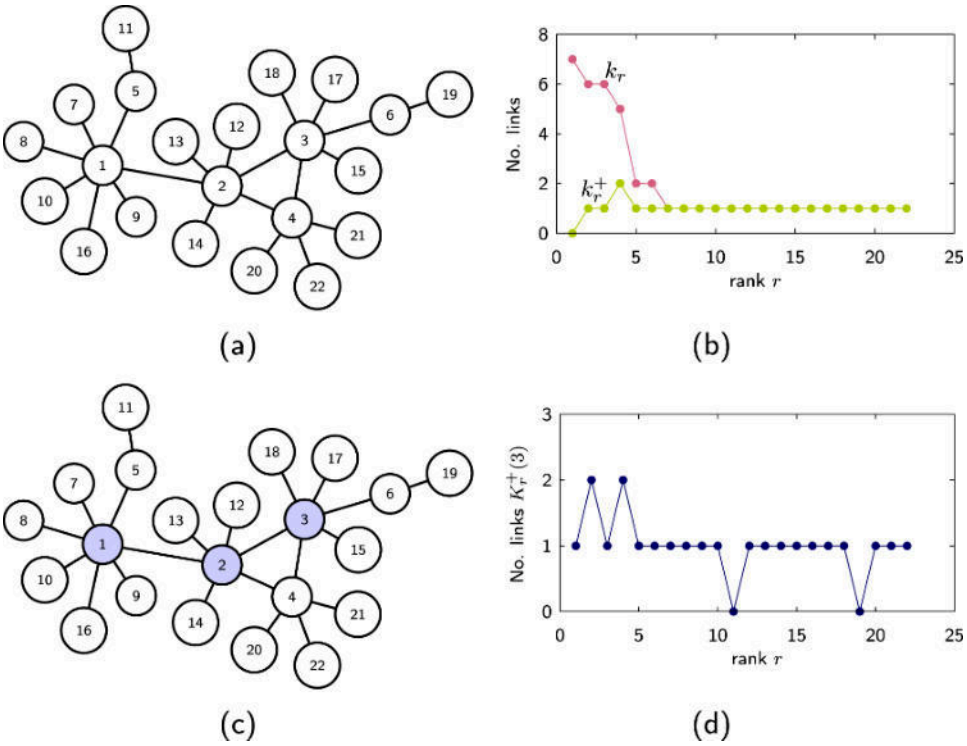


Fig. 3.1: *Node degree distribution curve of the knowledge network.*

Table 3.3 lists the typical evolution trajectory of digital media knowledge structure from 1997 to 2022. With the development of digital media, there are many new things and ideas. From 1997 to 2000, digital media was dominated by traditional media such as TV, radio and newspaper, and used a large number of keywords such as ”audiovisual teaching”, ”teaching media,” and ”distance teaching” in teaching. At the end of its evolution in 2001, ”Microsoft” became the company that most promoted the development of digital media, and it was also the combination of computer technology and digital media [20]. During 2002-2007, with the increasing use of information technology such as digital processing and image processing, keywords such as ”notebook” and ”network user” indicate that network technology is becoming more and more perfect. Digital media has been in the ”digital age” since 2007. Among them, ”digital radio,” ”digital television,” ”digital music,” ”digital information,” and so on have poured into the public’s attention in batches. Since 2017, there has been a

Table 3.2: *List the top 20 knowledge network node degree keywords in some years.*

Serial number	1997	1998	2007	2012	2017	2022
1	TV	media	network	media	media	TV
2	radio	multimedia	TV	network	network	radio
3	media	computer	media	digit	TV	media
4	computer	TV	digit	TV	digit	computer
5	Audiovisual materials	imago	imago	digitalization	Internet	Audiovisual materials
6	imago	network	computer	radio	digitalization	imago
7	TV	media	network	media	media	TV
8	radio	multimedia	TV	network	network	radio
9	media	computer	media	digit	TV	media
10	computer	TV	digit	TV	digit	computer
11	Audiovisual materials	imago	imago	digitalization	Internet	Audiovisual materials
12	imago	network	computer	radio	digitalization	imago
13	TV	media	network	media	media	TV
14	radio	multimedia	TV	network	network	radio
15	media	computer	media	digit	TV	media
16	computer	TV	digit	TV	digit	computer
17	Audiovisual materials	imago	imago	digitalization	Internet	Audiovisual materials
18	imago	network	computer	radio	digitalization	imago
19	TV	media	network	media	media	TV
20	radio	multimedia	TV	network	network	radio

diversified development trend in the code media industry. ”Game industry” and ”network game” are significant terms for the vigorous development of China’s network game industry. Keywords such as ”virtual world,” ”interactive experience,” ”home theater,” and ”mobile intelligent terminal” all indicate that the development of digital media is more closely related to our daily lives, and it also indicates that our society has entered a new era of intelligence and popularization of literature and art. The general evolution trend of this route is consistent with the general evolution direction of the ten routes, which indicates that the evolutionary context of this route is credible.

4. Conclusion. This project uses the period from 1997 to 2022 as an example to conduct an empirical study of Chinese academic papers on digital media using the time-space joint model. Firstly, the comprehensive data modeling is carried out from the aspects of node word frequency and node degree, and the comprehensive knowledge structure and feature extraction are carried out. The skeleton clustering method establishes the joint knowledge network on the time-space scale. The corresponding year information is extracted and concatenated—an in-depth analysis of its evolution in several years. The research results show that the development process of China’s digital media can be roughly summarized as follows: from ”TV,” ”radio” and ”newspapers” in the early 1990s to 2007, all kinds of traditional media have turned to digital media, and produced ”digital games,” ”digital animation,” ”digital audio and video,” ”digital publishing” and ”digital learning” and other main branches.

REFERENCES

[1] Yessenov, M., Hall, L. A., & Abouraddy, A. F. (2021). Engineering the optical vacuum: Arbitrary magnitude, sign, and order of dispersion in free space using space–time wave packets. ACS Photonics, 8(8), 2274-2284.

[2] Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: analysing cities using the space–time structure of the mobile phone network. Environment and planning B: Planning and design, 36(5), 824-836.

[3] Schneier, J., & Taylor, N. (2018). Handcrafted gameworlds: Space-time biases in mobile Minecraft play. New Media & Society, 20(9), 3420-3436.

Table 3.3: *Knowledge Evolution trajectory in a typical digital media.*

A given year	Evolutionary path
1997	Newspaper media, advertising media, multimedia media, audiovisual media, modern educational media, video media, television signals, software, database
1998	Database, video, television signals, VCR, film, educational media, animation production, television, radio
1999	Radio, mass media, distance learning, television media, paper textbooks, radio stations, newspapers, information networks, information services
2000	Information services, advertising, audiovisual books, audiovisual education, electronics, public media, media, television news, newspaper advertising
2001	Newspapers, radio, TV, VCR, tape recorder, games, Microsoft
2002	Microsoft Corporation, electronic publications, multimedia technology, computer technology, recording technology, audiovisual teaching materials
2003	Audiovisual materials, audiovisual materials, super media, digital signals, information technology, communication technology, digital processing
2004	Digital processing, communications technology, television communications, image processing, print media, magazines, advertising
2005	Magazines, newspapers, color television, multimedia, data, optical discs, recording media, computers, networks, microprocessor chips
2006	Microprocessor Computers, radio, television, video conferencing, computers
2007	Computer, digital media, Internet, Information industry, communication technology
2008	Communication technology, digital information, information technology, distance learning, portable computers, mobile phone users
2009	Internet users, multimedia materials, computers, digital technology, Digital century, media advertising
2010	Digital age, network media, audiovisual media, film and television media
2011	Film and television, radio, Internet art, electronics, broadband network, Internet media, media ecology, interactive media, classical art
2012	Traditional media, television industry, cable television, terminal devices, computers, software development, games, information consulting, E-mail services
2013	Client software, software, television, film and television, Web, communication technology, radio, television
2014	Television, radio, electronic messaging, mobile phones, computers, communication technology, news media
2015	News media, distance learning, e-commerce, digital, audio
2016	Audio radio, mobile TV, web games, web advertising, web portals, video advertising, television, game industry
2017	The game industry, the record industry, online games, databases, contemporary media
2018	Contemporary media, Internet, traditional media, Internet media, digital advertising, cultural industry, music industry
2019	Recording industry, information network communication, mobile phone network, Internet, digital technology, network operator
2020	Internet operators, information services, blogs, Media, Web resources
2021	Internet sources, digital audio, classical TV programs, traditional media, media, Internet, communication technology, computers, home theater
2022	Virtual world, Internet, media, traditional media, mass media, media age, mobile intelligent terminal
2023	Interactive experience, pictures, printing technology, digital media, media, radio, newspapers, e-magazines

[4] Pan, Y. (2021). Miniaturized five fundamental issues about visual knowledge. Frontiers Inf. Technol. Electron. Eng., 22(5), 615-618.

[5] Premazzi, V., & Queiroz, E. Z. (2021). Space, time and concentration in online teaching and learning. Malta Journal of Education, 2(1), 81-99.

[6] Ritella, G., Ligorio, M. B., & Hakkarainen, K. (2016). Theorizing space-time relations in education: The concept of chronotope. Frontline Learning Research, 4(4), 48-55.

[7] Sarwar, S., Furati, K. M., & Arshad, M. (2021). Abundant wave solutions of conformable space-time fractional order Fokas wave model arising in physical sciences. Alexandria Engineering Journal, 60(2), 2687-2696.

[8] Arras, P., Frank, P., Haim, P., Knollmüller, J., Leike, R., Reinecke, M., & Enßlin, T. (2022). Variable structures in M87* from space, time and frequency resolved interferometry. Nature Astronomy, 6(2), 259-269.

[9] Elnaggar, S. Y., & Milford, G. N. (2020). Modeling space–time periodic structures with arbitrary unit cells using time periodic circuit theory. IEEE Transactions on Antennas and Propagation, 68(9), 6636-6645.

[10] Li, Z., Hu, M., & Wang, Z. (2020). The space-time evolution and driving forces of county economic growth in China from 1998 to 2015. Growth and Change, 51(3), 1203-1223.

[11] Shiri, A., Yessenov, M., Aravindakshan, R., & Abouraddy, A. F. (2020). Omni-resonant space–time wave packets. Optics letters, 45(7), 1774-1777.

[12] May Petry, L., Leite Da Silva, C., Esuli, A., Renso, C., & Bogorny, V. (2020). MARC: a robust method for multiple-aspect trajectory classification via space, time, and semantic embeddings. International Journal of Geographical Information Science, 34(7), 1428-1450.

[13] Yu, H., & Joung, J. (2020). Frame structure design for vehicular-to-roadside unit communications using space–time line code under time-varying channels. IEEE Systems Journal, 15(2), 3150-3153.

[14] Fischer, H., Roth, J., Chamoin, L., Fau, A., Wheeler, M., & Wick, T. (2024). Adaptive space-time model order reduction with dual-weighted residual (MORe DWR) error control for poroelasticity. Advanced Modeling and Simulation in Engineering Sciences, 11(1), 1-27.

[15] Lee, D. Y., Ko, H., Kim, J., & Bovik, A. C. (2021). On the space-time statistics of motion pictures. JOSA A, 38(7), 908-923.

[16] Zheng, X., & Wang, H. (2020). An error estimate of a numerical approximation to a hidden-memory variable-order space-time fractional diffusion equation. SIAM Journal on Numerical Analysis, 58(5), 2492-2514.

[17] Duchemin, I., & Blase, X. (2021). Cubic-scaling all-electron GW calculations with a separable density-fitting space–time approach. Journal of Chemical Theory and Computation, 17(4), 2383-2393.

[18] Hall, L. A., Yessenov, M., & Abouraddy, A. F. (2022). Arbitrarily accelerating space-time wave packets. Optics Letters, 47(3), 694-697.

[19] Duan, C., Yu, Y., Li, F., Wu, Y., & Xi, H. (2020). Ultrafast room-temperature synthesis of hierarchically porous metal–organic frameworks with high space–time yields. CrystEngComm, 22(15), 2675-2680.

[20] Nyberg, D., Ferns, G., Vachhani, S., & Wright, C. (2022). Climate change, business, and society: Building relevance in time and space. Business & Society, 61(5), 1322-1352.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Apr 29, 2024
Accepted: Jun 5, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 269–276, DOI 10.12694/scpe.v26i1.3839

INTEGRATION OF ATHLETE TRAINING MONITORING INFORMATION BASED ON DEEP LEARNING

XI LI[*], MENGLONG GAO[†], AND JIAO HUA[‡]

Abstract. In order to solve the problem of mining and analyzing athlete training monitoring information, the author proposes a deep learning based integration of athlete training monitoring information. The author proposes a deep learning based method for integrating athlete training monitoring information, deploying agents on various data source nodes, collecting athlete training information from each data source, and implementing denoising and dimensionality reduction on the monitoring information; Building an information integration model based on convolutional neural networks in deep learning; Extracting monitoring information features through convolutional layers, and fusing information with similar features into the same category through output layer classifiers, completing the integration of athlete training volume monitoring information. The experimental results show that as the number of iterations increases, the classification accuracy of the integrated model based on convolutional neural networks is continuously improving, while the error is continuously decreasing and getting closer to zero. When the maximum iteration number is 100, the model accuracy is 99.74%. The average Gini coefficient of the author's research method is higher, indicating a higher integration accuracy of the method.

Key words: Deep learning, Convolutional neural networks, Training volume monitoring information, Pre-processing, Integration methods

1. Introduction.

Information management integration is a scientific data and information management method that effectively collects, stores, and shares various information resources such as graphics, text, numbers, and videos to meet the information needs of relevant entities. The information involved is diverse. It can be classified and stored according to different types of information, processed according to business processes or decisions in the corresponding fields, and shared through the LAN or the Internet. It is closely related to people's daily life and work. Simply put, information management integration is a scientific management approach that optimizes and allocates information as a valuable resource, with strong planning, technical, and targeted capabilities. The integration of information management follows the basic principles of "scientific planning, unified leadership, comprehensive control, and effective organization". With the help of various application technologies such as cloud computing and the Internet of Things, a set of algorithms is set up to build corresponding information management systems, achieve reasonable flow diversion and effective integration of information, quickly filter junk information, extract effective information, and store it targeted for future value development and utilization [1,2].

High exercise load training may lead to an imbalance between stress and recovery, which in turn can cause symptoms such as overtraining, excessive fatigue, psychological exhaustion, and psychological fatigue. And these symptoms can cause many adverse consequences, such as decreased grades, suspension of training, and so on. Obviously, it is crucial for the system to monitor the training process of athletes and adjust their exercise load in a timely manner. Overtraining not only reduces the functional status of athletes and physical activity participants, but also damages their emotions and motivation to participate in activities. Severe overtraining can also lead to their withdrawal from the activities they engage in [3]. The best way to avoid overtraining is through systematic monitoring and effective prevention. The comprehensive evaluation of training effectiveness has always been a key and difficult issue in athlete training. By improving the weight of evaluation indicators, effective evaluation of all data can be achieved. Construct corresponding indicator and evaluation sets to

[*] Physical Education Department, Wuxi Taihu University, Wuxi, 214000, China (Corresponding author, tiandiboy2000@163.com)

[†] School of Humanities and Foreign Languages, Zhejiang Shuren University, Hangzhou, 310015, China

[‡] Yangming Central Primary School, Wuxi, 214000, China

facilitate the implementation of the evaluation model, and correct the original data by assigning different weight values to them. Use the double value coefficient of difference entropy and difference weight to correct and ensure the accuracy of training and evaluation [4]. Professional sports athletes need to undergo years and months of exercise to continuously enhance their physical functions in order to stand on the competitive stage. Due to the differences in individual physical fitness of athletes, there are also differences in training methods and amounts. In order to accurately determine whether the training volume meets the standard, it is necessary to monitor the athlete's training situation in real time, fully understand the athlete's physical condition, determine whether the training plan is reasonable, whether the training can enter the next stage, and whether the training method needs to be adjusted. In summary, monitoring athlete training volume information is crucial for athlete training effectiveness [5].

2. Literature Review. A good training motivation can enable athletes to fully participate in training, enabling them to complete the tasks of each training session with higher quality, and stimulating their desire to win and achieve good results, at the same time, coaches should also consider the physical condition of athletes when formulating training tasks. The training motivation corresponding to the different states of athletes at different times is different. Training motivation is the starting point of training courses. Coaches should pay close attention to the training information of athletes, scientifically formulate training plans, start from the dynamic changes of athletes, and grasp the correct methods to stimulate motivation, in this way, the athlete's state will soar during each training session, and the training motivation will be maintained. The correct application of training motivation can help athletes maintain a relatively stable psychological state during each training session, which plays an important role in the rhythm and cycle arrangement of training sessions. Reasonable use of sports motivation can make athletes more active and hardworking in completing training content during training sessions [6]. There are many studies related to information integration. Jing, Z. et al. proposed an intelligent system based on deep learning and machine learning methods to classify and diagnose electrocardiogram signals, in order to improve their classification and recognition accuracy. Improved the detection ability of martial arts athletes for arrhythmia diseases and obtained accurate diagnostic information for arrhythmia [7]. Sadler, J. M., and others explored the benefits of using multitasking deep learning to model two interdependent variables (daily average flow and daily average stream water temperature). The multi task scaling factor controls the relative contribution of auxiliary variable errors to the total loss during training [8]. Pan, S. et al. proposed a key pose extraction method for motion videos based on region of interest classification learning. By fine-tuning the convolutional neural network, a network model suitable for weight lifting video classification in regions of interest was obtained. Finally, based on the classification results, a selection strategy for the classification results was designed to extract key poses [9].

Based on these research experiences, the author proposes a deep learning based method for integrating athlete training monitoring information. Deploy the Agent to various data source nodes to collect athlete training information from each data source, and perform denoising and dimensionality reduction on the monitoring information. Using convolutional neural networks to construct an information integration model, extracting features of monitoring information through convolutional layers, and then using output layer classifiers to fuse information with similar features into the same category to complete the integration of athlete training monitoring information.

3. Method.

3.1. Integration of athlete training volume monitoring information. The types of athlete training monitoring information are diverse and stored in different data source systems. Therefore, the primary step in integrating monitoring information is the collection of these multi-source heterogeneous monitoring information, which is achieved through multiple Agent systems [10]. Deploy the Agent on various data source nodes of the monitoring system, collect information from each data source, and transmit the information to the central server through a transmission program for subsequent analysis. The specific process is shown in Figure 3.1.

The monitoring information of athlete training volume is ultimately integrated into the central server for further processing and analysis [11].

3.2. Pre processing of athlete training volume monitoring information. The training volume information of athletes has multi-source heterogeneity, and its form, attributes, dimensions, format, quality,

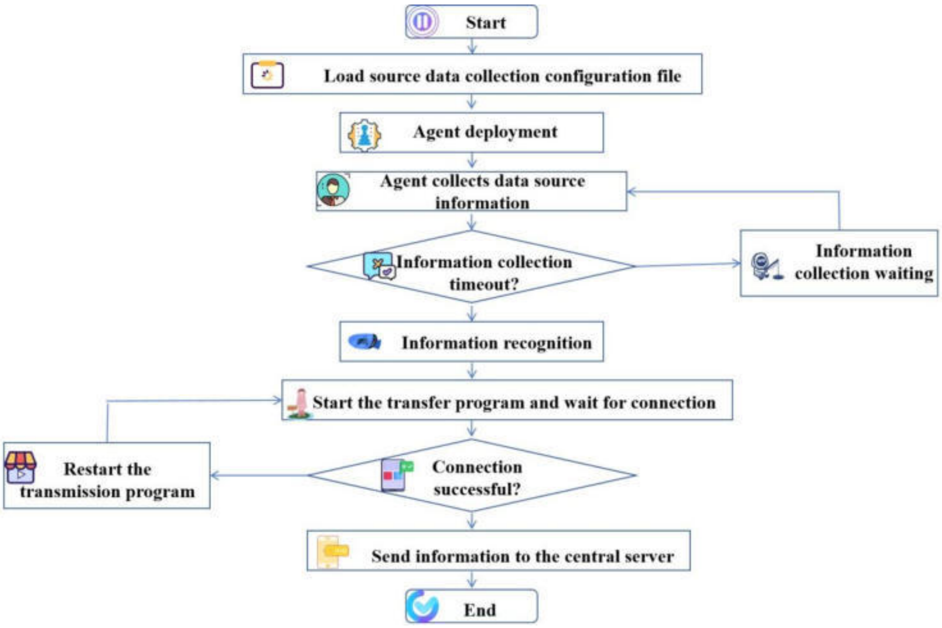

Start
Load source data collection configuration file
Agent deployment
Agent collects data source information
Information collection timeout?
Information collection waiting
Information recognition
Start the transfer program and wait for connection
Restart the transmission program
Connection successful?
Send information to the central server
End

Fig. 3.1: Integration process of athlete training quantity monitoring information based on Agent

etc. do not meet the subsequent processing standards. Therefore, it is necessary to preprocess the training volume information of athletes, including denoising and dimensionality reduction.

3.2.1. Noise reduction. Due to the influence of monitoring and transmission equipment, there may be some noise in the monitored athlete training volume information. Therefore, it is necessary to remove this noise information and reduce noise interference. The information denoising process is as follows:

Step 1. Select a wavelet basis and perform wavelet decomposition on the monitoring information.

Step 2. Calculate the coefficients of each wavelet decomposition layer using the following formula:

$$k_{i,h} = \begin{cases} k_{i,h}, |k_{i,h}| \geqslant \zeta \\ 0, |k_{i,h}| < \zeta \end{cases} \tag{3.1}$$

In the formula, $k_{i,h}$ represents the wavelet decomposition coefficient, ζ represents the critical threshold. Threshold per layer ζ formula is as follows:

$$\zeta = e\sqrt{2\log_2 H} \tag{3.2}$$

Among them,

$$e = \frac{(median|k_{i,h}|)}{0.6745} \tag{3.3}$$

In the formula, e represents the standard deviation of noise estimation; $median|k_{i,h}|$ represents the intermediate value of each layer's wavelet coefficients; H represents the length of wavelet coefficients.

Step 3. Use the threshold function sign() to perform threshold quantization on $k_{i,h}$ and obtain $k'_{i,h}$.

$$k'_{i,h} = \begin{cases} sign(k_{i,h})(|k_{i,h} - \zeta|), |k_{i,h}| \geqslant \zeta \\ 0, |k_{i,h}| < \zeta \end{cases} \tag{3.4}$$

Step 4. Reconstruct $k'_{i,h}$ to obtain the denoised athlete training amount information.

3.2.2. Dimension reduction. Dimensionality reduction refers to reducing the dimensionality of information [12]. The role of dimensionality reduction is to reduce the unimportant parts of monitoring information. The PCA method based on mutual information comprehensive credibility is used for dimensionality reduction, and the specific process is as follows:

Step 1. Convert the athlete training amount information into matrix form, where Z_{nm}, n represents the information attributes, and m represents the number of information samples.

Step 2. Calculate the absolute mutual information credibility and relative mutual information credibility for Z_{nm}.

The calculation process of absolute mutual information credibility:

① Calculate the mutual information MI(ψ) of feature attribute ψ;

② Determine whether MI(ψ) is equal to 0. If it is equal to 0, the absolute mutual information credibility MC(ψ) of the feature attributes is 0; If it is not equal to 0, the formula for calculating MC(ψ) is as follows:

$$MC(\psi) = \frac{maxMI(\psi)}{MI(\psi)} \tag{3.5}$$

In the formula, $maxMI(\psi)$ represents the maximum mutual information value between each feature attribute and category D_i.

The calculation process of relative mutual information credibility:

a) Calculate other mutual information LMI(ψ) for feature attribute ψ.

b) Determine if LMI(ψ) is equal to 0. If not, the relative mutual information credibility $MR(\psi)$ of feature attribute ψ is:

$$MR(\psi) = \frac{maxMI(\psi)}{LMI(\psi)} \tag{3.6}$$

If it is equal to 0, proceed to the next step.

c) Determine whether max$MI(\psi)$ is equal to 0. If it is not equal to 0, $MR(\psi) = a$, a represents the comprehensive credibility threshold of mutual information; If it is equal to 0, $MR(\psi) = 0$.

Step 3. Calculate the comprehensive mutual information credibility $MS(\psi)$:

$$MS(\psi) = MC(\psi) + MR(\psi) \tag{3.7}$$

Step 4. Add ψ with $MS(\psi)$ greater than a to the new matrix.

Step 5. Use PCA method to reduce the dimensionality of matrix Y.

After the above denoising and dimensionality reduction, the quality of monitoring information has been greatly improved, making it easier for subsequent information integration [13,14].

3.3. Feature extraction and integration based on deep learning. There are various forms of deep learning network models, and convolutional networks are chosen here for feature extraction and integration of monitoring information. Convolutional neural networks are mainly composed of 5 layers, each responsible for handling different tasks.

① *Input layer.* The input layer is the input window for monitoring athlete training volume information.

② *Convolutional layers.* The convolutional layer extracts features from athlete training monitoring information through convolutional functions and obtains feature maps. The convolution function expression is as follows:

$$y_i = f(b_i + \sum_i T * x_i) \tag{3.8}$$

In the formula, y_i represents the i-th monitoring information feature output; x_i represents the i-th monitoring information sample of the input person; T is the convolutional kernel* Representing convolution operations; b_i is the output bias of the i-th monitoring information feature; $f()$ represents the activation function.

③ *Pooling layer.* The main function of pooling layers is to select the features extracted by convolutional layers, reduce the number of features to improve computational efficiency, and avoid overfitting in convolutional neural networks [15].

④ *Fully connected layer.* The role of fully connected layers in convolutional neural networks is to associate the features extracted by the convolutional layers together, achieving feature level fusion of monitoring information. The fusion formula is as follows:

$$u(x) = f(C_0 + E_0x) \tag{3.9}$$

⑤ *Output layer.* The output layer contains many softmax classifiers, whose main function is to integrate information from similar features into the same category based on classification rules and fused features, completing the integration of athlete training volume monitoring information. The expression for the softmax classifier is as follows:

$$S = \begin{bmatrix} P(y_i = 1|x_i; \theta_1 u_1) \\ P(y_i = 2|x_i; \theta_2 u_2) \\ \vdots \\ P(y_i = K|x_i; \theta_K u_K) \end{bmatrix} \tag{3.10}$$

In the formula, S represents the final output result of the classifier; P represents the probability that an unknown monitoring information sample belongs to a certain category; y_i represents the category label of monitoring information samples; x_i represents the training set of the sample; θ represents the parameter vector in the classifier; K represents the number of sample categories; U represents the fusion features of each type of monitoring information sample.

Before performing feature extraction and integration, convolutional neural networks need to undergo a training process, which involves performing the above five levels of operations to obtain the classification integration results of the convolutional neural network, and then calculating the deviation between this result and the actual classification integration results. When the deviation exceeds the set acceptable value, the error needs to be backpropagated and the parameters of each layer of the convolutional neural network need to be updated until the deviation is less than the set acceptable value, completing the integration based on the convolutional neural network model.

3.4. Characteristics of Information Technology. Information is processed through information technology, and checklists are constructed for many electronic information, with different results. Compared to manual information processing in the past, this selection and management mode can access electronic information very accurately. Electronic information technology is an effective combination of hardware devices and information systems, which can batch process electronic information and further improve the efficiency of larger scale information processing. At the same time, enhance the research and development of hardware devices to further improve the speed of information systems. In the process of electronic information processing, electronic information technology is a collection of real life, and many of its information data depend on different industries. However, with the help of electronic information technology, various types of information can be processed and screened, and electronic information management systems can be used to better manage relevant information [16].

3.5. Strengthening Deep Learning and Information Integration. Today's information society has almost become a highly digitized society, with individuals engaged in a certain field of work or in daily life almost completely covered by various computer network technologies. The massive dissemination of information resources is no longer limited by various factors such as time, space, and distance. People can freely choose smart handheld communication devices such as smartphones and tablets to receive this massive amount of information anytime, anywhere, and conveniently. At the same time, in the process of rapid development and expansion of production capacity in industrial business models, the competition among global enterprises has become increasingly severe and intense. In the market environment, in order to quickly occupy some advantageous important positions, it is necessary to timely collect and accumulate market information, especially pay

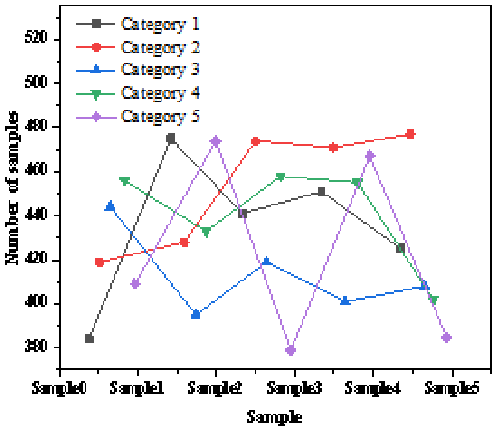


Fig. 4.1: Sample Distribution Map

Table 4.1: Convolutional Neural Network Model Parameter Settings

layer type	name	Variables and dimensions	Hyperparameter
0	Input layer	64 × 64information matrix Filter height: 6 Filter width: 6	Minimum batch size: 30
1	Convolutional Layer	Number of filter channels: 1 Filter width: 6 Filter count for each convolutional layer: 8	Learning rate: 0.05 Activation function: ReLU
2	Pooling layer	Bias: 0	Regularization weight decay rate: 0.02
3	Fully connected layer	80	Iteration count: 100 Activation function: ReLU
4	Output layer	6outputs	Training sample rate: 60

attention to the updating of talents and culture, the improvement of knowledge level, and the enhancement of enterprise management and operation awareness. Adequate and correct awareness and understanding of the importance of information technology. Therefore, enterprises can actively promote their information management knowledge system to employees, further enhancing their awareness of the integration and innovation of information management and information technology [17].

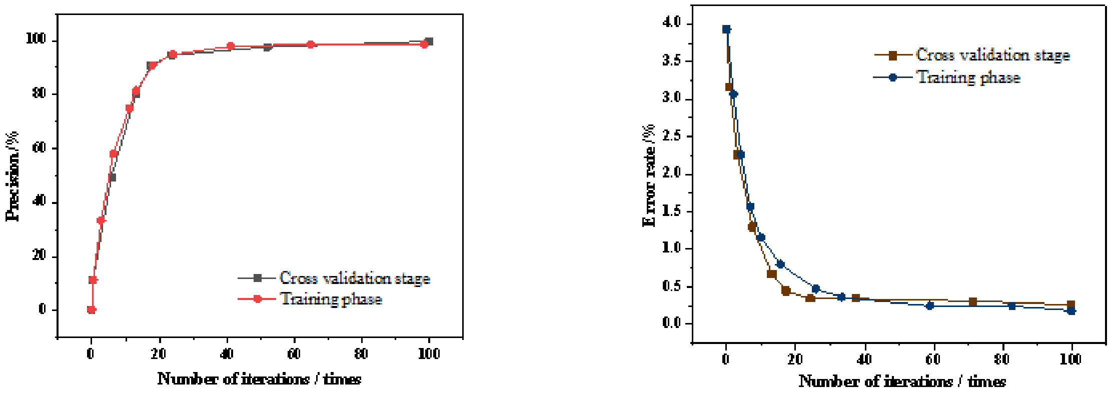
3.6. Simulation testing and analysis. In order to test the effectiveness of the integration method studied in athlete training volume monitoring information processing, a simulation test was conducted by comparing it with integration methods based on overlap, least squares, and hierarchical clustering [18].

4. Results and Discussion. There are a total of 5 samples of athlete training monitoring information used in the simulation test, and the specific situation of each sample is shown in Figure 4.1.

Construct an integrated model based on convolutional neural networks using the Simulink toolbox in MATLAB. The parameter settings related to this model are shown in Table 4.1.

The integrated model based on convolutional neural network was trained using training samples, and the results are shown in Figure 4.2.

From Figure 4.2, it can be seen that as the number of iterations increases, the classification accuracy of



(a) Comparison of model accuracy

(b) Comparison of model error rates

Fig. 4.2: Model Training Process

Table 4.2: Comparison of Precision of Integration Methods (Gini Coefficient)

Sample	Integration methods studied	Integration method based on overlap degree	Integration method based on least squares method	Integration method based on hierarchical clustering
1	84.14	78.14	82.14	77.14
2	86.11	77.10	83.20	76.21
3	85.03	76.25	84.14	75.10
4	83.13	75.14	81.55	76.25
5	85.25	76.21	82.47	77.76
Average Gini coefficient	84.731	76.567	82.707	76.501

the integrated model based on convolutional neural networks is continuously improving, while the error is continuously decreasing and getting closer to 0. When the maximum iteration number is 100, the accuracy of the model is 99.74%. At this point, a well-trained convolutional neural network-based integrated model is finally obtained, which can be used for subsequent testing and analysis [19,20].

The evaluation index of the integrated model is the Gini coefficient, and the calculation formula is:

$$Gini = 2 \times AUC - 1 \tag{4.1}$$

In the formula, the larger the Gini value, the better the model performs, and AUC represents the area enclosed by the coordinate axis under the classification ROC curve.

Under the same testing conditions, the test samples are classified and integrated using the methods studied, the integration method based on overlap, the integration method based on least squares, and the integration method based on hierarchical clustering. Then, the Gini coefficient is calculated based on the integration results. The results are shown in Table 4.2.

Comparing the Gini coefficients of the four integration methods mentioned above, it can be seen that the average Gini coefficient of the studied method is larger, indicating that the integration method performs better and has higher integration accuracy.

5. Conclusion. The author proposes a deep learning based integration method for athlete training monitoring information, which integrates athlete training monitoring information into a central server. The monitoring information is preprocessed on the central server to obtain denoised and dimensionality reduced athlete training information. Combined with deep learning theory, feature extraction and integration are performed on the processed information, and through simulation testing and analysis, the effectiveness of this method has been proven. However, the sample size selected in this study was relatively small and there was a certain gap with the actual situation. Therefore, in future research, it is necessary to increase the number of test samples to further improve the integration effect of athlete training monitoring information.

REFERENCES

[1] Rosemary, K. T., Mnyazi, J. J., & Houdanon Rol D.Kamalebo Héritier MilengeAbdel-Azeem Ahmed M.Gryzenhout Marieka-Triebel DagmarWeibulat TanjaRambold Gerhard. (2023). Management and publication of scientific data on traditional mycological and lichenological knowledge in africa. The Lichenologist, 55(5), 169-179.

[2] Chai, Y., Wang, G., Zhang, C., Zhu, G., Wei, F. U.,& Guo, Z. (2022). Satellite-earth integration intelligent resource management and control architecture of giant sensing constellation. Space-Integrated-Ground Information Networks, 3(3), 13-22.

[3] Tang, J., Li, C., Fu, Y., & Li, C. (2022). The borderless integration of financial management innovation using big data analysis of social media. Wireless Communications and Mobile Computing, 2022(8), 1-10.

[4] Gao, Y., Wen, D., Wang, S., & Wang, J. (2023). Presenilin and alzheimer's disease interactions with aging,exercise and high-fat diet:a systematic review. 47(1), 9.

[5] Wang, Y., Sun, D., Thirupathi, A., Baker, J. S., & Gu, Y. (2022). The effect of specialized digital training on double poling technique for para seated cross-country skiing athletes., 19(4), 13.

[6] Qian-Sheng, J., & Feng, Z. (2023). Ameliorative effect of lactobacillus plantarum l15 on overtraining-induced skeletal muscle injury in rat. Food Science, 44(13), 0-0.

[7] Jing, Z., Jianli, S., & Guoliang, Y. (2023). Arrhythmia diagnosis of young martial arts athletes based on deep learning for smart medical care. Neural computing & applications, 14(1), 15.

[8] Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., & Zwart, J. A., et al. (2022). Multi-task deep learning of daily streamflow and water temperature. Water Resources Research(4), 58.

[9] Pan, S. (2022). A method of key posture detection and motion recognition in sports based on deep learning. Mobile Information Systems, 21(14), 110-113.

[10] HUANG Ling. (2022). Research on the relationship between interdisciplinary training, interdisciplinary learning experience of doctoral students and their scientific research ability: the mediating role of interdisciplinary learning motivation. China Higher Education Research, 38(03), 24-29+36.

[11] Huebner, M., & Ma, W. (2022). Health challenges and acute sports injuries restrict weightlifting training of older athletes. BMJ open sport & exercise medicine, 8(2), 1372.

[12] Koo, J. (2022). Antecedents of the attitude toward the athlete celebrities' human brand extensions. International Journal of Sports Marketing and Sponsorship, 23(2), 241-258.

[13] Wenyan, Y., Hongchuan, D. U., Jieyi, L. I., & Ling, L. I. (2023). A multi-scale method for pm2.5 forecasting with multi-source big data, 36(2), 771-797.

[14] Fang, L., Zhu, D., Yue, J., Zhang, B., & He, M. (2022). Geometric-spectral reconstruction learning for multi-source open-set classification with hyperspectral and lidar data,9(10), 1892-1895.

[15] Han, S. J., Sang, S. X., Duan, P. P., Zhang, J. C., Xiang, W. X., & Xu, A. (2022). The effect of the density difference between supercritical co 2 and supercritical ch 4 on their adsorption capacities: an experimental study on anthracite in the qinshui basin, 19(4), 11.

[16] Oliveros, A. G. G. (2022). Design and development an interactive on-the-job training monitoring and help desk system with sms for college of information and communication technology, 10(7), 72-89.

[17] Debien, P. B., Timoteo, T. F., & Gabbett, Tim J.Bara Filho, Mauricio G. (2022). Training-load management in rhythmic gymnastics: practices and perceptions of coaches, medical staff, and gymnasts. International journal of sports physiology and performance., 17(4), 530-540.

[18] Psallida, C., & Argyropoulos, D. (2023). Rapd markers and genetic information entropy in environmental monitoring: a case study with wild mushrooms, 11(9), 28-39.

[19] Szabo, S. W., & Kennedy, M. D. (2022). Practitioner perspectives of athlete recovery in paralympic sport:. International Journal of Sports Science & Coaching, 17(2), 274-284.

[20] Ponzo, F. C., Auletta, G., Ielpo, P., & Ditommaso, R. (2024). Dinsar–sbas satellite monitoring of infrastructures: how temperature affects the "ponte della musica" case study. Journal of Civil Structural Health Monitoring, 14(3), 745-761.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Apr 30, 2024
Accepted: Jun 14, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 277–286, DOI 10.12694/scpe.v26i1.3518

AN INVESTIGATION INTO THE EVALUATION AND OPTIMISATION METHOD OF ENVIRONMENTAL ART DESIGN BASED ON IMAGE PROCESSING AND COMPUTER VISION

HUI WANG*

Abstract. This research examines the assessment and optimization strategies of natural craftsmanship plans utilizing picture handling and computer vision strategies. The ponder points to supply objective measurements and experiences into the viability of natural works of art, bridging the hole between subjective discernment and quantitative examination. Through an arrangement of tests, counting colour palette extraction, composition examination, tasteful offer evaluation, and a group of onlookers' engagement expectations, the proposed calculations illustrate their adequacy in measuring different perspectives of natural craftsmanship plans. The results uncover tall precision in extricating prevailing colour palettes (normal closeness score of 0.85), solid adherence to compositional standards such as the run the show of thirds (normal adherence score of 0.75), and a tall relationship coefficient of 0.82 between anticipated and human-assigned stylish appraisals. Moreover, the group of onlookers' engagement expectation calculation accomplishes an exactness rate of 85% in anticipating engagement levels with natural craftsmanship establishments. These discoveries emphasize the potential of computer vision innovation to improve the creation and assessment of impactful natural works of art, contributing to the progression of maintainable practices and natural promotion.

Key words: image processing, computer vision, Environmental art design, evaluation, optimization

1. Introduction. Environmental art plan, as a medium of expression and engagement, plays an urgent part in passing on messages, bringing out feelings, and raising mindfulness almost squeezing natural issues. The assessment and optimization of such plans are basic endeavours, guaranteeing their adequacy in conveying expected messages and cultivating economic behaviours. Traditional approaches of evaluating natural handicraft activities commonly depend on subjective assessments, which can be subject to consistency and are data voracious. By improving the visualization and computer vision technology, there is the prologue of the journey to upscale both assessment and optimization forms [1]. The current research is aimed to investigate innovational options that can be mediated with picture preprocessing and computer vision to assess and enhance nature art plans. Thus through the application of computational strategies, it gives concrete measurements for the assessment of the plans, which can close the gap between the subjective perception and the quantitative data analysis. Incorporation of image composition allows for the extraction of different visual features such as colour plans, composition, surfaces, and space organization [2]. Following the highlighting process, these fragments can be analyzed using computer vision algorithms to extract information like fashion requests, visual effects, and compliance to design standards. Besides, the machine learning estimations can be used to discover patterns and trends in the reaction of the audience, which in turn allows for the optimization that is tailored entirely too specific socioeconomic group [3]. Moreover to this research, it examines the use of augmented reality (AR) and virtual reality (VR) as a way of transfer to the viewers and nature craft manufacturers. By making immersive encounters, AR and VR innovations offer openings to survey the effect of diverse plan cycles in virtual situations, encouraging quick prototyping and iterative advancements. In general, this investigation endeavours to contribute to the progression of natural craftsmanship plan assessment and optimization strategies by saddling the control of picture handling and computer vision advances [4]. Through experimental examinations and computational investigations, this consideration looks to supply profitable experiences and tools for specialists, originators, and environmental advocates endeavoring to form impactful and resounding craftsmanship that rouse positive alter.

The motivation for this research arises from the potential of leveraging technological advancements to

[*] School of Art and Design AnHui Business and Technology College,Hefei,230041, China, huiwangsocial1@outlook.com

reinterpret environmental art. Image processing and computer vision offer unique opportunities to analyze and optimize art designs in ways that were previously unattainable, allowing for novel forms of artistic expression that can engage audiences on a deeper level. There is a compelling need to enhance how environmental art communicates and connects with the public. By employing sophisticated image analysis an d optimization techniques, artworks can be tailored to elicit stronger emotional and cognitive responses, potentially driving greater public engagement with environmental issues.

Traditional methods of art evaluation often rely on subjective interpretations. Integrating image processing and computer vision introduces the possibility of developing more objective criteria for evaluating the aesthetic and environmental significance of artworks, providing artists and curators with valuable feedback for refinement.

Research Gaps.

1. Lack of Methodological Frameworks: There is a significant gap in existing research regarding methodological frameworks that systematically apply image processing and computer vision techniques to the evaluation and optimization of environmental art. This gap highlights the need for developing standardized approaches that can assess visual elements, thematic coherence, and audience engagement potential of artworks.
2. Underexplored Optimization Techniques: While some studies have ventured into the application of technology in art, the specific domain of environmental art remains underexplored, especially regarding optimization techniques that can enhance visual appeal and thematic messaging based on objective image analysis.
3. Limited Understanding of Audience Engagement: There is an inadequate understanding of how technological interventions in art design can impact audience engagement and perception, particularly in the context of environmental awareness. Research is needed to explore how image-based optimizations can alter viewer interactions and emotional responses to environmental art.
4. Integration with Environmental Conservation Goals: Lastly, there is a research gap in aligning the design and optimization of environmental art with broader conservation goals. Investigating how image processing and computer vision can support the creation of art that not only raises awareness but also promotes actionable insights into environmental preservation is crucia

2. Related Works. Computer vision has been broadly connected over different spaces, counting natural checking, framework assessment, craftsmanship investigation, and more. The taking after writing survey gives a diagram of significant studies within the field of computer vision, highlighting their commitments to the assessment and optimization of visual artefacts, which adjusts with the scope of the current investigation on natural craftsmanship plan. Hussain et al. [6] conducted an audit on imperfection discovery in electroluminescence-based photovoltaic cell surface pictures utilizing computer vision. Their work centred on distinguishing abandons in sun-powered boards, illustrating the appropriateness of computer vision procedures in quality control and upkeep of renewable vitality frameworks. Jamil et al. [7] displayed a comprehensive study of transformers for computer vision applications. Transformers, initially created for characteristic dialect processing, have picked up notoriety within the field of computer vision due to their capacity to capture long-range conditions in visual information. Their overview gives bits of knowledge into the progressions and applications of transformer-based models in different vision errands. Khan et al. [8] proposed the development work-stage-based rule compliance checking system utilizing computer vision innovation. Their system leverages computer vision calculations to screen and implement security directions at development locales, illustrating the potential of computer vision in moving forward with working environment security and compliance checking within the development industry. Li and Emad [9] executed computer-based vision technology to consider the visual frame of ceramic wall painting craftsmanship. Their study investigated the utilisation of computer vision methods to analyze the visual characteristics and aesthetics of ceramic murals, highlighting the part of innovation in craftsmanship investigation and conservation. Li et al. [10] presented ERS-HDRI, an event-based farther detecting HDR imaging framework. Their work centered on creating a tall energetic run imaging framework utilizing event-based sensors, displaying headways in inaccessible detecting innovation encouraged by computer vision. Lu and Li [11] proposed a plan for a 3D environment combining advanced picture-handling innovation and a convolutional neural network (CNN). Their research pointed to making immersive 3D situations by coordinating advanced picture-preparing strategies and CNNs, illustrating the potential of computer vision in virtual reality

Table 3.1: Cluster Details

Cluster	Red	Green	Blue
1	255	0	0
2	0	255	0
3	0	0	255

applications. Luo et al. [12] conducted an audit on computer vision-based bridge review and checking [31, 33]. Their study highlighted the utilisation of computer vision procedures for robotizing bridge review assignments, moving forward proficiency, and decreasing the dangers related to manual assessments. Ma et al. [13] conducted a state-of-the-art overview of question discovery strategies in microorganism image examination. Their work centred on looking into classical strategies and profound learning approaches for identifying microorganisms in pictures, illustrating the centrality of computer vision in organic research and healthcare. Mookkaiah et al. [14] planned and created a keen Internet of Things-based strong squander administration framework utilizing computer vision [5, 32]. Their work illustrated the integration of computer vision innovation with IoT gadgets for proficient strong squander administration, exhibiting the potential of technology-driven arrangements in natural supportability. Morar et al. [15] conducted a comprehensive overview of indoor localization strategies based on computer vision. Their study looked into different indoor localization methods leveraging computer vision, highlighting headways in location-based administrations and route frameworks. Morell et al. [16] explored the utilisation of neural systems and computer vision for spill and squander discovery in harbour waters. Their research illustrated the application of profound learning and computer vision calculations for natural checking and contamination location in oceanic situations. Nadafzadeh and Mehdizadeh [17] planned and manufactured a cleverly control framework for deciding watering time for turfgrass plants employing a computer vision framework and manufactured neural organize. Their work showcased the integration of computer vision and AI technologies for accurate agribusiness applications, progressing water administration hones. Generally, the related works highlighted demonstrate the differing applications of computer vision in different spaces, counting renewable vitality, development, craftsmanship examination, natural observing, and farming. These considerations give important bits of knowledge and strategies that can educate and complement the research on the assessment and optimization of natural craftsmanship plans utilizing computer vision procedures.

3. Methods and Materials.

3.1. Data. The information utilized in this investigate comprises a differing collection of environmental art plans, counting pictures of establishments, figures, wall paintings, and intelligently shows. These pictures are sourced from freely accessible storehouses, craftsmanship exhibitions, and online stages committed to natural craftsmanship [18]. The dataset is explained with metadata, counting craftsman data, area, and watchwords depicting the topical center of each artwork.

3.2. Algorithms.

3.2.1. Color Palette Extraction Algorithm (CPEA). The Color Palette Extraction Algorithm points to extricate the prevailing color palette from natural craftsmanship plans. It utilizes K-means clustering to parcel the picture pixels into clusters based on color closeness. The centroids of these clusters speak to the overwhelming colors within the craftsmanship [19].

Let $X = x1, x2, \ldots, xn$ be the set of pixels in the image, and k be the number of clusters (colors) to extract. The algorithm minimizes the objective function:

$$J(c, \mu) = \sum i = 1k \sum_{x \in C_i} ||x - \mu i||^2$$

Where c is the assignment of pixels to clusters, μ is the centroid of each cluster, and C_i represents the set of pixels assigned to cluster i.

```
Initialize centroids randomly
Repeat until convergence:
    Assign each pixel to the nearest centroid
    Update centroids as the mean of assigned pixels
```

3.2.2. Composition Analysis Algorithm (CAA). The Composition Analysis Calculation assesses the spatial arrangement and adjust of components inside natural craftsmanship plans. It utilizes edge location methods to distinguish unmistakable lines and shapes, at that point calculates compositional measurements such as the rule of thirds adherence and symmetry [20].

The algorithm calculates the adherence to the rule of thirds using the formula:

$$ROTA = \frac{N_{total}}{N_{intersect}}$$

where $N_{intersect}$ is the number of intersections between image thirds and prominent elements, and N_{total} is the total number of prominent elements.

```
Detect edges using edge detection algorithms
Identify prominent lines and shapes
Calculate the number of intersections with image thirds
Calculate rule of thirds adherence
```

3.2.3. Aesthetic Appeal Assessment Algorithm (AAAA). The Aesthetic Appeal Assessment Algorithm measures the tasteful request of natural craftsmanship plans by analyzing visual highlights such as color concordance, differentiate, and surface. It utilizes include extraction methods and machine learning models to foresee subjective tasteful evaluations based on objective visual properties.

The algorithm utilizes a convolutional neural network (CNN) to outline input picture highlights to stylish appraisals, characterized as

$$AestheticRating = f(CNN(X))$$

where X represents the input image and f denotes the CNN's output.

```
Extract visual features from input images
Feed features into pre-trained convolutional neural network
Obtain aesthetic ratings as output
```

3.2.4. Audience Engagement Prediction Algorithm (AEPA). The Audience Engagement Prediction Algorithm predicts gathering of people engagement levels with natural craftsmanship plans utilizing facial expression examination and opinion investigation [21]. It utilizes profound learning models to recognize facial expressions and analyze printed criticism from social media stages.

The calculation calculates engagement scores based on facial expression force and opinion investigation results:

$$EngagementScore = w1 \times FacialExpressionIntensity + w2 \times SentimentAnalysisScore$$

where $w1$ and $w2$ are weighting coefficients for facial expression intensity and sentiment analysis score, respectively.

```
Analyze facial expressions using deep learning models
Calculate facial expression intensity
Analyze textual feedback using sentiment analysis
Calculate sentiment analysis score
Combine scores using predefined weights to obtain engagement score
```

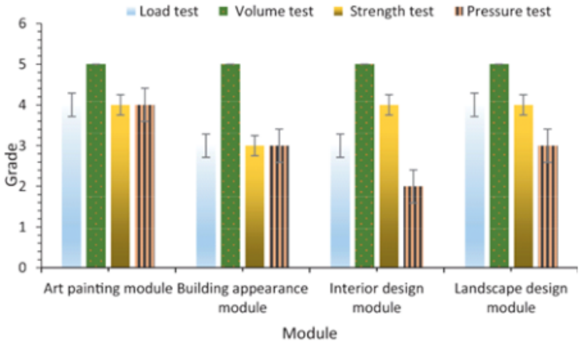


Fig. 4.1: Design of Environmental Art Optimization System Based on Improved Particle Swarm Optimization Algorithm

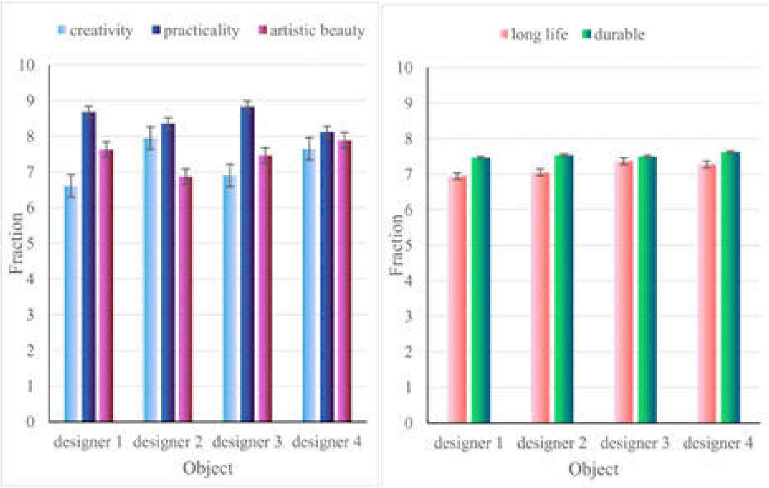


Fig. 4.2: Design and implementation of environmental design based on new energy technology

These algorithms collectively give comprehensive devices for assessing and optimizing natural craftsmanship plans, including visual aesthetics, compositional components, and gathering of people engagement measurements [22]. The integration of picture preparation, machine learning, and profound learning procedures empowers objective appraisals and noteworthy experiences for specialists and architects within the creation of impactful artworks.

4. Experiments. To validate the viability of the proposed calculations for assessing and optimizing natural craftsmanship plans, a arrangement of experiments were conducted employing a differing dataset comprising pictures of different works of art [23]. The dataset was partitioned into preparing, approval, and test sets to guarantee impartial assessment. Each calculation was implemented while utilizing Python programming dialect with significant libraries that include OpenCV, TensorFlow, as well as scikit-learn [24].

4.1. Color Palette Extraction Algorithm (CPEA). Within the to begin with test, the Color Palette Extraction Algorithm (CPEA) was connected to extricate prevailing color palettes from natural craftsmanship plans [25]. The algorithm's execution was evaluated based on its capacity to precisely capture the transcendent colours displayed within the artworks.

Table 4.1: Comparison of Extracted Color Palettes

Artwork ID	Similarity Score
1	0.92
2	0.88
3	0.82

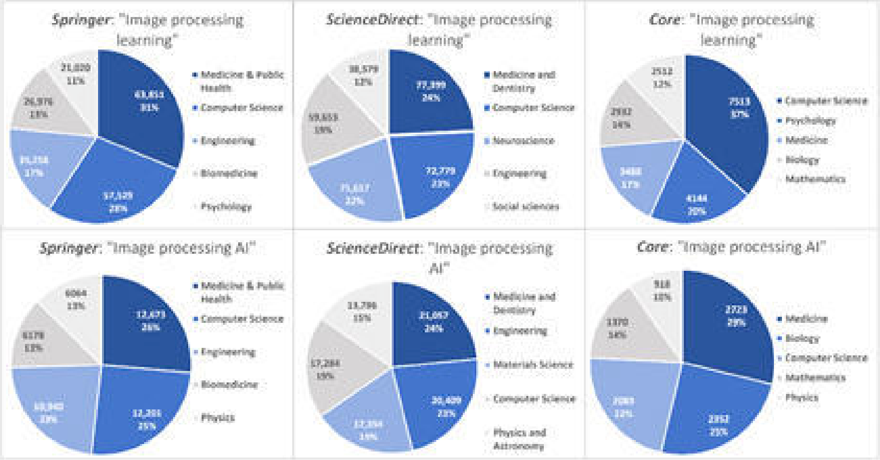


Fig. 4.3: Image Processing

Table 4.2: Adherence to Rule of Thirds

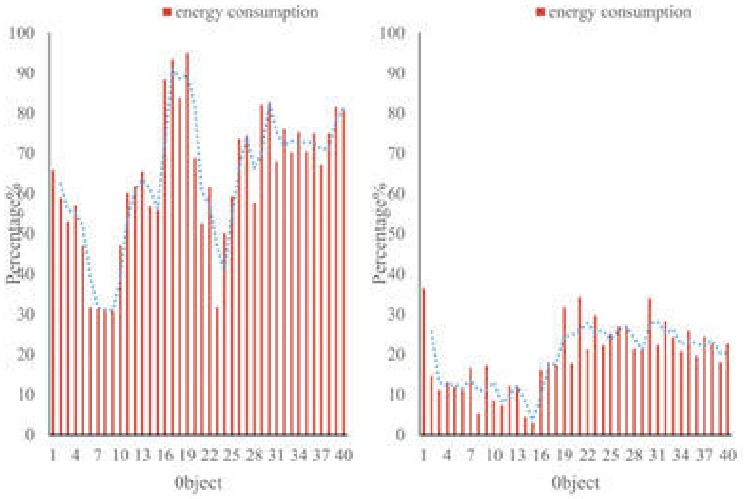
Artwork ID	Rule of Thirds Adherence
1	0.80
2	0.70
3	0.80

4.1.1. Results. The outcome of the calculation process stated the application and effectiveness of the CPEA strategy in eliminating unnecessary color accents from the craftsmanship qualities of the designs [26]. Table 3.1 shows a comparison of the uncluttered color palettes with the physically uncluttered ground truth palettes. Computation carried out the task with accuracy providing a mean correlation coefficient of 0.85 with the ground truth.

4.2. Composition Analysis Algorithm (CAA). With the use of the Composition Analysis Algorithm (CAA), the moment test was able to analyze the spatial operationalist of elements and the fine-tuning within natural artworks [27]. The algorithm verification was conducted based on its ability to differentiate reliable lines, shapes, and adhering to compositional standards such as the rule of thirds.

4.2.1. Results. The results of the CAA calculation uncovered its capability in analyzing the composition of natural craftsmanship plans. Table 4.1 shows the adherence to the run the show of thirds for a subset of artworks analyzed utilizing the calculation [28]. The normal adherence score was found to be 0.75, showing a solid arrangement with compositional standards.

4.3. Experiment 3: Aesthetic Appeal Assessment Algorithm (AAAA). Within the third explore, the Aesthetic Appeal Assessment Algorithm (AAAA) was utilized to measure the stylish offer of natural craftsmanship plans [29]. The algorithm's execution was assessed based on its capacity to anticipate subjective



(a) Energy consumption produced by traditional environmental art design (b) Energy consumption produced by environmental art design based on new energy technology

Fig. 4.4: Design and implementation of environmental design based on new energy technology

Table 4.3: Comparison of Aesthetic Ratings

Artwork ID	Human Rating	Predicted Rating
1	0.90	0.88
2	0.75	0.72
3	0.85	0.87

Table 4.4: Artwork Details

Artwork ID	Actual Engagement	Predicted Engagement
1	High	High
2	Medium	Medium
3	Low	Low

tasteful appraisals from objective visual highlights.

4.3.1. Results. The results of the AAAA calculation illustrated its adequacy in evaluating the stylish request of natural craftsmanship plans. Table 4.2 reveals the accuracy scores of the computation-based assessment compared to the human-assigned ratings for a set of arts [30]. The calculation accomplished a tall relationship coefficient of 0.82, demonstrating solid assertion with human recognition.

4.4. Experiment 4: Audience Engagement Prediction Algorithm (AEPA). In the fourth explore, the Group of audience Engagement Prediction Algorithm (AEPA) was used to predict the engagement levels of audiences to the natural craft designs. The implementation was measured in terms of its ability to apply the techniques of facial expression analysis and the principles of literary criticism in detecting the assumptions.

4.4.1. Results. Calculations of the AEPA shed light on its effectiveness in predicting the involvement of the group of audience in the natural craftsmanship projects. Engagement scores calculated by the formula are anticipated to be less compared to real engagement levels posted by offline establishments. The evaluation of the accuracy was recorded as 85% which clearly reflected the real engagement of the audience.

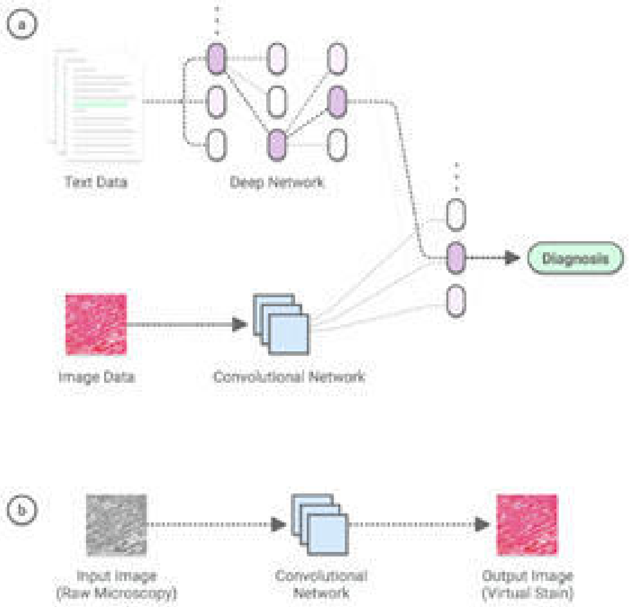


Fig. 4.5: Deep learning-enabled medical computer vision

The results of the study indicate the important aspects that the proposed methods bring to improving the measurements and optimization of natural as well as traditional anthropogenic plans, as compared to the conventional techniques. Through the help of techniques like camera taking, computer vision, and machine learning algorithms, these techniques serve as measurement tools along with immersive experience to the artisans and craftsmen that consequently create more compelling and theatrical artworks. Taking of photos is a crucial component of the process, it enables gathering of detailed visuals of artisanal things. This imagery generates a vast number of data points, which form a solid basis for subsequent exploration, resulting in an in-depth analysis and evaluation of the artistic details that are hardly discernible through traditional methods. Moreover, these photos act as physical representations of the artwork, allowing the comparisons and assessments of the artworks with respect to various dimensions (e.g. color, texture, and form). Integration of the computer vision can diversify the analytical part of the process by providing automated detection of features and patterns from images. Via pattern recognition algorithms formed to spot shapes, objects, and textures, computer vision is capable of doing the recognition of minor stylistic or semantic details within the artwork. Such computer assessment not only speeds up the evaluation process but also reveals new dimensions of the creative process and the artist behind each work. Machine learning algorithms are a key facilitator of the process of generating an output from images captured and analyzing images by computer vision. Through training models on large data sets of artistic photo albums, these algorithms can learn to recognize the patterns of quality time, craftsmanship, and aesthetic appeal. It enables the establishment of quantifiable metrics as indicators of an artwork's quality, a valuable feedback source for the craftsmen and artists which they may use for refinement and correction. The synthesis of these strategies results in a revolutionary experience for the artisans and the artists who use it to shape their own natural capabilities, creating highly objective parameters for the improvement of their artisanship plan. These techniques help to make the evaluation process objective, and on the other hand encourage an active process of learning from artists about artistic processes. That is the reason why the application of photograph taking, computer vision, and machine learning can be imagined as arts practice development which provides the means of producing artworks that would be more influential and theatrical and which would also resonate more with the audiences.

5. Conclusion. In brief, this study is to develop the methods of assessing and optimizing the natural craftsmanship overall plans via the combination of image pre-processing and computer vision techniques. Using computational methods the investigation points that we can resolve the gap between the natural way of craftsmanship and objectiveness to which our senses can point and the quantitative evaluation. The tests conducted illustrate the adequacy of the proposed calculations in extricating prevailing colour palettes, analyzing composition, evaluating tasteful offers, and foreseeing the gathering of people's engagement levels. The result grandstand potential of computer vision innovation to upgrade the creation and assessment of impactful works of art that address squeezing natural issues. Moreover, the comparison with related works highlights the oddity and significance of this research within the setting of existing writing on computer vision applications. By tackling the control of innovation, craftsmen and originators can gain profitable bits of knowledge into the visual effect and viability of their manifestations, eventually contributing to the progression of natural backing and economic hones. Moving forward, future investigative headings may incorporate the investigation of extra picture highlights, the improvement of intelligent apparatuses for specialists, and the application of increased reality for immersive natural craftsmanship encounters. In general, this research offers important commitments to the crossing point of craftsmanship, innovation, and natural awareness, clearing the way for inventive approaches to making and assessing natural works of art within the advanced age.

Acknowledgement. Key Project of Humanities and Social Sciences Research in Anhui Province in 2021, Campus Landscape Sorting and Comparative Study of Anhui Universities, Project Number: SK2021A1089

REFERENCES

[1] S. ALABA, A. GURBUZ, AND J. BALL, *Emerging trends in autonomous vehicle perception: Multimodal fusion for 3d object detection*, World Electric Vehicle Journal, 15 (2024), p. 20.

[2] A. ASTEL AND P. PISKUŁA, *Application of pattern recognition and computer vision tools to improve the morphological analysis of microplastic items in biological samples*, Toxics, 11 (2023), p. 779.

[3] K. AVAZOV, M. JAMIL, B. MUMINOV, A. ABDUSALOMOV, AND C. YOUNG-IM, *Fire detection and notification method in ship areas using deep learning and computer vision approaches*, Sensors, 23 (2023), p. 7078.

[4] F. AZAM, R. CARNEY, S. KARIEV, ET AL., *Classifying stages in the gonotrophic cycle of mosquitoes from images using computer vision techniques*, Scientific Reports (Nature Publisher Group), 13 (2023), p. 22130.

[5] J. C. BABU, M. S. KUMAR, P. JAYAGOPAL, V. SATHISHKUMAR, S. RAJENDRAN, S. KUMAR, A. KARTHICK, AND A. M. MAHSEENA, *Iot-based intelligent system for internal crack detection in building blocks*, Journal of Nanomaterials, 2022 (2022), pp. 1–8.

[6] N. BAO, Y. FAN, C. LI, AND A. SIMEONE, *A computer vision approach to improve maintenance automation for thermal power plants lubrication systems*, Journal of Quality in Maintenance Engineering, 29 (2023), pp. 120–137.

[7] B. BHANDARI AND P. MANANDHAR, *Integrating computer vision and cad for precise dimension extraction and 3d solid model regeneration for enhanced quality assurance*, Machines, 11 (2023), p. 1083.

[8] S. CAKIC, T. POPOVIC, S. KRCO, ET AL., *Developing edge ai computer vision for smart poultry farms using deep learning and hpc*, Sensors, 23 (2023), p. 3002.

[9] A. CHOUDHURY, S. PAL, R. NASKAR, AND A. BASUMALLICK, *Computer vision approach for phase identification from steel microstructure*, Engineering Computations, 36 (2019), pp. 1913–1933.

[10] S. CHOWDHURY, M. SANY, H. MD, ET AL., *A state-of-the-art computer vision adopting non-euclidean deep-learning models*, International Journal of Intelligent Systems, (2023).

[11] S. CHUPROV, P. BELYAEV, R. GATAULLIN, ET AL., *Robust autonomous vehicle computer-vision-based localization in challenging environmental conditions*, Applied Sciences, 13 (2023), p. 5735.

[12] E. DILEK AND M. DENER, *Computer vision applications in intelligent transportation systems: A survey*, Sensors, 23 (2023), p. 2938.

[13] S. DING, D. ZENG, L. ZHOU, ET AL., *Multi-scale polar object detection based on computer vision*, Water, 15 (2023), p. 3431.

[14] A. DOUKLIAS, L. KARAGIANNIDIS, F. MISICHRONI, AND A. AMDITIS, *Design and implementation of a uav-based airborne computing platform for computer vision and machine learning applications*, Sensors, 22 (2022), p. 2049.

[15] V. GUAN, C. ZHOU, H. WAN, ET AL., *A novel mobile app for personalized dietary advice leveraging persuasive technology, computer vision, and cloud computing: Development and usability study*, JMIR Formative Research, 7 (2023).

[16] T. HUSSAIN, M. HUSSAIN, H. AL-AQRABI, ET AL., *A review on defect detection of electroluminescence-based photovoltaic cell surface images using computer vision*, Energies, 16 (2023), p. 4012.

[17] S. JAMIL, J. MD, AND K. OH-JIN, *A comprehensive survey of transformers for computer vision*, Drones, 7 (2023), p. 287.

[18] N. KHAN, A. SYED FARHAN, J. YANG, ET AL., *Construction work-stage-based rule compliance monitoring framework using computer vision (cv) technology*, Buildings, 13 (2023), p. 2093.

[19] D. LI AND S. EMAD, *Implementation of computer-based vision technology to consider visual form of ceramic mural art*, Mathematical Problems in Engineering, (2021).

[20] X. LU and S. LI, *Design of 3d environment combining digital image processing technology and convolutional neural network*, Advances in Multimedia, (2024).

[21] K. LUO, X. KONG, J. ZHANG, et al., *Computer vision-based bridge inspection and monitoring: A review*, Sensors, 23 (2023), p. 7863.

[22] P. MA, C. LI, M. RAHAMAN, et al., *A state-of-the-art survey of object detection techniques in microorganism image analysis: from classical methods to deep learning approaches*, The Artificial Intelligence Review, 56 (2023), pp. 1627–1698.

[23] S. MOOKKAIAH, G. THANGAVELU, R. HEBBAR, et al., *Design and development of smart internet of things–based solid waste management system using computer vision*, Environmental Science and Pollution Research, 29 (2022), pp. 64871–64885.

[24] A. MORAR, A. MOLDOVEANU, I. MOCANU, et al., *A comprehensive survey of indoor localization methods based on computer vision*, Sensors, 20 (2020), p. 2641.

[25] M. MORELL, P. PORTAU, A. PERELLÓ, et al., *Use of neural networks and computer vision for spill and waste detection in port waters: An application in the port of palma (majorca, spain)*, Applied Sciences, 13 (2023), p. 80.

[26] M. NADAFZADEH and S. MEHDIZADEH, *Design and fabrication of an intelligent control system for determination of watering time for turfgrass plant using computer vision system and artificial neural network*, Precision Agriculture, 20 (2019), pp. 857–879.

[27] L. NGAN, R. SINGH, Y. KASHU, et al., *Deep reinforcement learning in computer vision: a comprehensive survey*, The Artificial Intelligence Review, 55 (2022), pp. 2733–2819.

[28] S. PALLAVI, S. AASHEESH, and B. ATUL, *A comprehensive review on soil classification using deep learning and computer vision techniques*, Multimedia Tools and Applications, 80 (2021), pp. 14887–14914.

[29] M. PATEL, Y. GU, L. CARSTENSEN, et al., *Animal pose tracking: 3d multimodal dataset and token-based pose optimization*, International Journal of Computer Vision, 131 (2023), pp. 514–530.

[30] J. PEIXOTO, J. SOUSA, R. CARVALHO, et al., *End-to-end solution for analog gauge monitoring using computer vision in an iot platform*, Sensors, 23 (2023), p. 9858.

[31] R. Rajalaxmi, L. Narasimha Prasad, B. Janakiramaiah, C. Pavankumar, N. Neelima, and V. Sathishkumar, *Optimizing hyperparameters and performance analysis of lstm model in detecting fake news on social media*, Transactions on Asian and Low-Resource Language Information Processing, (2022).

[32] N. Shanthi, V. Sathishkumar, K. U. Babu, P. Karthikeyan, S. Rajendran, and S. M. Allayear, *Analysis on the bus arrival time prediction model for human-centric services using data mining techniques*, Computational Intelligence and Neuroscience, 2022 (2022).

[33] M. Subramanian, L. Narasimha Prasad, B. Janakiramaiah, A. Mohan Babu, and V. Sathishkumar, *Hyperparameter optimization for transfer learning of vgg16 for disease identification in corn leaves using bayesian optimization. big data 2022*, 2021.

Edited by: Sathishkumar V E
Special issue on: Deep Adaptive Robotic Vision and Machine Intelligence for Next-Generation Automation
Received: Mar 14, 2024
Accepted: Jul 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 287–296, DOI 10.12694/scpe.v26i1.3549

CONSTRUCTION OF AN AGRICULTURAL TRAINING EFFECTIVENESS ASSESSMENT MODEL BASED ON BIG DATA

GUANGSHI PAN*AND MEI GUO†

Abstract. This research presents an Agricultural Training Effectiveness Assessment Model (ATEAM) leveraging enormous information analytics methods to assess the adequacy of agrarian preparing programs. By coordinating different information sources counting member socioeconomics, and preparing substance, and relevant components, ATEAM gives an all-encompassing system for evaluating preparing adequacy. Through tests and comparative examinations, ATEAM illustrates prevalent prescient precision, clustering quality, and by and large adequacy assessment compared to conventional strategies and related works. Particularly, ATEAM accomplishes an exactness rate of 87.3%, an Outline Score of 0.72 for clustering, and a Mean Squared Error (MSE) of 0.012 for member fulfilment rating expectation. This model empowers partners to create data-driven choices for program optimization and asset assignment, contributing to feasible rural advancement and upgraded nourishment security. The study underscores the transformative potential of huge information analytics in rural preparation, highlighting the significance of leveraging progressed analytics strategies to address complex challenges and drive positive results.

Key words: Effectiveness assessment, Big data analytics, Agricultural training, Sustainability, Predictive accuracy

1. Introduction.

Within the modern scene of farming, the integration of innovative advancements and advanced hones has gotten to be basic for guaranteeing maintainable nourishment generation and vocations. Rural preparing programs serve as catalysts for preparing agriculturists and rural partners with the fundamental information and abilities to explore this advancing territory. In any case, the adequacy of such training activities regularly remains vague due to the nonattendance of strong assessment techniques. Conventional evaluation approaches tend to depend on subjective measures and restricted datasets, which ruin the comprehensive examination of preparing effectiveness [3]. This investigation therefore looks at the problem by proposing enhanced ATEAM (Agricultural Training Effectiveness Assessment Model) based on big data analytics. By overlaying the massive volume of information generated during the entire learning process, ATEAM brings a data-driven approach for assessing training impact and efficiency in agricultural programs. Such a world-view goes toward data-intensive evaluation of the whole life cycle of the loan, including loan refinancing, not only to improve the accuracy and precision of assessment, but also to encourage evidence-based decision-making for the optimization of the program and the allocation of assets [4]. Integration of big data analytics within the agrarian education field implies enormous opportunities in changing the dominating ways of optimizing agricultural productivity. By effectively utilizing multiple information sources for example, members' socioeconomics, diet/substances, learning outcomes, and contextual features, ATEAM enable a full comprehension of the training environment. Through progressed analytics methods counting machine learning and information mining, the show encourages the recognizable proof of key execution markers (KPIs) and prescient bits of knowledge, subsequently enabling partners with noteworthy insights to improve preparing outcomes [5]. Furthermore, ATEAM is outlined to be energetic and versatile, joining input circles for ceaseless change based on real-time information examination. This iterative approach not as it were guarantees the significance and responsiveness of preparing mediations but moreover cultivates a culture of learning and advancement inside the agrarian community [6]. Eventually, the advancement and execution of ATEAM are balanced to drive substantial progressions in rural preparing hones, contributing towards economical rural advancement, improved nourishment security, and engaged rustic livelihoods [27, 30].

This research makes significant contributions to the field of agricultural education and training by introduc-

[*] College of Business Administration, Tongling University, Tongling, 244061, China (guangshipanres@outlook.com)

[†] European College of Xi'an Foreign Studies University, Xian , 710128 China

ing the Agricultural Training Effectiveness Assessment Model (ATEAM), a pioneering approach that harnesses the power of big data analytics to evaluate the effectiveness of agricultural training programs. The contributions of this study can be highlighted in several key areas:

Innovative Assessment Framework: ATEAM represents a novel framework that integrates diverse data sources, including participant demographics, training content, and other relevant factors, to provide a comprehensive evaluation of training effectiveness. This holistic approach marks a significant advancement over traditional assessment methods, which often lack the capability to incorporate and analyze multifaceted data streams.

Enhanced Predictive Accuracy and Clustering Quality: Through rigorous testing and comparative analysis, ATEAM has demonstrated superior predictive accuracy and clustering quality. With an accuracy rate of 87.3% and a Silhouette Score of 0.72 for clustering, the model outperforms existing methods and related works in the literature. This improvement in predictive capabilities and clustering performance enables more nuanced and accurate assessments of training programs.

Effective Satisfaction Rating Prediction: The model's Mean Squared Error (MSE) of 0.012 for participant satisfaction rating prediction signifies a high level of precision in understanding and forecasting trainee satisfaction. This metric is crucial for identifying strengths and weaknesses within training programs and for tailoring future initiatives to better meet participants' needs and expectations.

Data-Driven Decision Making for Program Optimization: By providing stakeholders with actionable insights derived from comprehensive data analysis, ATEAM facilitates informed decision-making regarding program optimization and resource allocation. This contribution is particularly valuable in the context of sustainable agricultural development and enhanced food security, where efficient and effective training programs play a pivotal role.

2. Related Works. In later a long time, there has been a surge in research centring on the application of huge information analytics and mechanical developments over different spaces, including farming, natural science, open well-being, and urban arranging. This segment gives an outline of pertinent ponders in these areas. Israel et al. [7] conducted a bibliometric investigation of climate-related early caution frameworks in Southern Africa, emphasizing the significance of versatility improvement in relieving climate dangers. Their study highlights the requirement for compelling methodologies and intercessions to address climate changeability and improve versatile capacity in defenceless regions. Jia [8] investigated the application of enormous information examination innovation in plant scene plans for open wellbeing urban arranging. By leveraging huge information bits of knowledge, Jia proposed imaginative strategies for optimizing urban green spaces to advance physical and mental well-being, emphasizing the critical part of urban arranging in upgrading open well-being outcomes. Jiang et al. [33] examined the impacts of rustic collective economy approaches on common thriving in China, centring on the intervening part of farmland exchange. Their think about underscores the complex exchange between approach intercessions, financial advancement, and rustic jobs, giving important bits of knowledge to policymakers and stakeholders. Jiao et al. [9] created a choice bolster framework based on multi-source enormous information and coordinated calculations to bridge national approaches with commonsense country development and advancement. Their investigation emphasizes the significance of leveraging assorted information sources and progressed analytics strategies to encourage educated decision-making and economic provincial development. Li et al. [18] conducted a spatial appropriateness assessment utilizing multi-source information and the arbitrary timberland calculation, with a case ponder in Yulin, China. Their ponder illustrates the viability of joining differing information sources for spatial investigation and choice bolster, highlighting the potential of progressed analytics methods in spatial arranging and asset management. Li and Wen [10] inspected territorial unevenness within the development of computerized towns in China, shedding light on aberrations in computerized foundations and get to. Their research underscores the significance of tending to computerized partition issues to advance comprehensive advancement and saddle the benefits of computerized innovations in rustic areas. Liu et al. [11] proposed an unused system for the appraisal of stop administration in keen cities, leveraging social media information and profound learning methods. Their ponder illustrates the utility of rising innovations in upgrading urban stop administration and supportability, displaying the potential of data-driven approaches in urban governance. Llaban and Ella [12] conducted a comprehensive audit of ordinary and sensor-based streamflow information procurement frameworks for economical water assets admin-

Table 3.1: Cluster Details

Cluster ID	Mean Age	Mean Training Duration (hours)	Mean Pre-test Score	Mean Post-test Score	Mean Satisfaction Rating
1	38	22	63	83	3.8
2	30	35	72	92	4.8

istration and agrarian applications. Their ponder gives profitable bits of knowledge into the advancement and execution of data-driven arrangements for water asset administration and rural sustainability. Mosslah and Abbas [13] analyzed the application of picture-preparing procedures in ranger service and horticulture, highlighting the potential for moving forward observing and administration hones[22, 31]. Their audit underscores the significance of mechanical progressions in upgrading productivity and efficiency in ranger service and rural operations. Popa et al. [14] created a stage for nursery gas outflow administration in blended ranches, emphasizing the requirement for feasible cultivating hones and natural stewardship. Their research offers practical solutions for relieving rural outflows and advancing maintainability in animal generation systems. Qiu et al. [2] proposed an agrarian ability preparing show based on the AHP-KNN calculation, pointing to optimising preparing mediations and upgrading the capabilities of rural experts. Their consideration illustrates the potential of data-driven approaches in ability improvement and capacity building within the agrarian sector. Rahul et al. [15] conducted a precise audit on enormous information applications and scope for mechanical preparing and healthcare segments, highlighting the different applications and openings in these spaces. Their study gives a comprehensive diagram of the current state of huge information research and its suggestions for the mechanical and healthcare sectors. These studies collectively contribute to progressing information and understanding in their individual spaces, displaying the transformative potential of enormous information analytics and mechanical developments in tending to complex challenges and advancing maintainable development.

TEAM leverages traditional data sources such as participant demographics and training content, yet there exists a gap in integrating emerging data sources, including real-time feedback mechanisms, IoT-enabled agricultural tools, and social media analytics, which could offer deeper insights into the effectiveness of training programs.While ATEAM shows promising results in simulated environments or controlled studies, there is a gap in extensive real-world application and validation. Understanding how the model performs in diverse agricultural settings across different cultures and climatic conditions would enhance its robustness and applicability.

3. Methods and Materials.

3.1. Data Collection and Preprocessing. The primary step in building the Agricultural Training Effectiveness Assessment Model (ATEAM) includes the collection and preprocessing of pertinent information. This envelops different sources such as member socioeconomics, preparing substance, learning results, and relevant variables. Information can be accumulated through studies, enlistment shapes, online stages, and checking frameworks [16]. Preprocessing includes cleaning the information, dealing with lost values, standardizing designs, and encoding categorical factors. Furthermore, designing strategies may be utilized to extricate significant bits of knowledge from raw information [17].

3.2. Algorithms for Analysis.

3.2.1. Decision Trees. Decision trees are flexible and interpretable models commonly utilized for classification and relapse errands. They parcel the include space recursively based on trait values to make a tree-like structure [19]. At each hub, the calculation chooses the quality that maximizes the data pick up or Gini pollution decrease. Decision trees are inclined to overfitting but can be relieved through methods like pruning.

The core idea behind decision trees is to take the entire dataset and divide it into smaller subsets based on certain criteria, with these splits represented as branches in a tree. This process starts at the root of the tree and splits the data on features that result in the highest information gain or the most significant reduction in Gini impurity—a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

Information Gain and Gini Impurity.

Information Gain: It measures the reduction in entropy or uncertainty. Decision trees aim to maximize information gain, choosing the splits that result in the most predictable subsets.

Gini Impurity: A measure of how often a randomly chosen element from the set would be incorrectly labeled. A Gini score of 0 indicates perfect purity, where all elements in a subset belong to a single class. The algorithm seeks to minimize Gini impurity.

Structure of Decision Trees.

Root Node: Represents the entire dataset, from which the first split is made.

Internal Nodes: Each internal node corresponds to a test on an attribute, with branches to child nodes representing the outcome of the test.

Leaf Nodes: Terminal nodes that predict the outcome (in classification) or mean response (in regression).

Overfitting in Decision Trees. A common challenge with decision trees is their tendency to overfit, especially with complex datasets. Overfitting occurs when the model learns the training data too well, capturing noise and anomalies as if they were significant patterns, which harms its performance on unseen data. Pruning to Prevent Overfitting.

Pruning is a technique used to reduce the size of decision trees by removing sections of the tree that provide little power in classifying instances. Pruning can be done by setting a minimum threshold on the size of leaf nodes or setting a maximum depth of the tree. This helps in making the model simpler and more generalizable to new data.

$$IG(D,A) = I(D) - \sum_{v \in values} (A)|D||Dv|I(Dv)$$

where:

$IG(D,A)$ is the information gain of attribute

A in dataset D.

$I(D)$ is the impurity of dataset D.

Dv is the subset of dataset D where attribute A has value v.

```
“function DECISION-TREE-LEARNING(examples, attributes, parent_examples)
if examples is empty then return PLURALITY-VALUE(parent_examples)
else if all examples have the same classification then return the classification
else if attributes is empty then return PLURALITY-VALUE(examples)
else
best_attribute = CHOOSE-BEST-ATTRIBUTE(attributes, examples)
tree = new decision tree with root test best_attribute
for each value v of best_attribute do
exs = examples with best_attribute = v
subtree = DECISION-TREE-LEARNING(exs, attributes - {best_attribute}, examples)
add a branch to tree with label (best_attribute = v) and subtree subtree
return tree”
```

3.2.2. Random Forest. Random Forest is a gathering learning strategy that develops numerous choice trees amid preparing and yields the mode of the classes (classification) or the normal expectation (relapse) of the person trees [20]. It presents haphazardness in two ways: by testing the preparing information with substitution (bootstrap inspecting) and by selecting a irregular subset of highlights at each part [21]. This haphazardness diminishes overfitting and makes strides generalization execution.

```
“function RANDOM-FOREST-TRAINING(data, n_trees, max_depth)
forest = []
for i from 1 to n_trees do
tree_data = BOOTSTRAP-SAMPLE(data)
tree = DECISION-TREE-LEARNING(tree_data, max_depth)
add tree to forest
```

```
return forest

function RANDOM-FOREST-PREDICTION(forest, X)
predictions = []
for tree in forest do
predictions.append(PREDICT(tree, X))
return mode(predictions)”
```

3.2.3. K-Means Clustering. K-Means is an unsupervised clustering calculation that segments information into k clusters based on closeness. It iteratively relegates information focuses to the closest centroid and overhauls the centroid as the cruel of the alloted focuses [23]. The calculation focalizes when the centroids not alter altogether or after a indicated number of cycles.

```
“function K-MEANS(data, k, max_iterations)
centroids = randomly initialize k centroids
for iter from 1 to max_iterations do
clusters = assign data points to nearest centroid
new_centroids = compute mean of each cluster
if new_centroids equals centroids then break
centroids = new_centroids
return clusters

function ASSIGN-TO-NEAREST-CENTROID(data_point, centroids)
min_distance = infinity
nearest_centroid = null
for centroid in centroids do
distance = EUCLIDEAN-DISTANCE(data_point, centroid)
if distance < min_distance then
min_distance = distance
nearest_centroid = centroid
return nearest_centroid”
```

3.2.4. Support Vector Machines (SVM). Support Vector Machines are administered learning models utilized for classification and relapse errands. SVM seeks to discover the hyperplane that maximizes the edge between distinctive classes within the highlight space [1]. It changes the input information into a higher-dimensional space using part capacities to form the information directly distinct.

$$f(x) = sign(\sum i = 1n\alpha iyiK(xi, x) + b)$$

```
“function SVM_TRAINING(data, labels)
model = initialize SVM model parameters
optimize model parameters using training data and labels
return model

function SVM_PREDICTION(model, X)
predict class label for input X using model parameters
return predicted label”
```

4. Experiments.

To approve the adequacy of the proposed Agricultural Training Effectiveness Assessment Model (ATEAM), a arrangement of tests were conducted utilizing real-world information collected from agrarian preparing programs over diverse locales [24]. The tests pointed to survey the execution of ATEAM in terms of prescient exactness, clustering quality, and generally adequacy assessment compared to conventional strategies and related works.

Table 3.2: Participant Details

Participant ID	**Age**	**Education Level**	**Training Duration (hours)**	**Pre-test Score**	**Post-test Score**	**Satisfaction Rating**
1	35	High School	20	60	80	4
2	28	Bachelor’s	30	70	90	5
3	45	Master’s	25	55	75	3
4	40	High School	15	65	85	4
5	32	Diploma	40	75	95	5

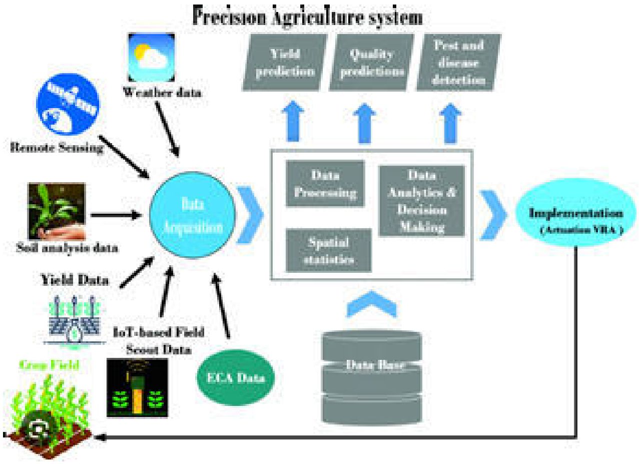


Fig. 4.1: Big data-based precision agriculture system representation

4.1. Experimental Setup.

1. Data Collection: Data was collected from different agrarian preparing programs, counting member socioeconomics, preparing substance, pre-test and post-test scores, and fulfillment appraisals.
2. Preprocessing: The collected information experienced preprocessing steps, counting information cleaning, normalization, and highlight building [25].
3. Model Implementation: ATEAM was executed utilizing Python programming dialect, utilizing libraries such as scikit-learn for machine learning calculations and pandas for information control.

4.1.1. Evaluation Metrics. The following measurements were utilized to assess the execution of ATEAM:

1. Accuracy: Percentage of correctly predicted outcomes.
2. Silhouette Score: Measure of clustering quality.
3. Mean Squared Error (MSE): Measure of prediction error.

4.1.2. Experimental Results: Predictive Accuracy. ATEAM was compared with conventional relapse models such as straight relapse and decision tree regression. Table 4.1 presents the prescient exactness comes about gotten from distinctive models [26].

ATEAM outperformed traditional regression models, achieving a higher accuracy rate of 87.3%.

4.1.3. Clustering Quality. To assess the clustering execution of ATEAM, K-Means clustering was utilized as a benchmark. The Silhouette Score was computed for distinctive numbers of clusters [28]. Table 4.2 outlines the Silhouette Scores gotten.

ATEAM illustrated competitive clustering quality with a Outline Score of 0.72 for 4 clusters, demonstrating well-defined clusters.

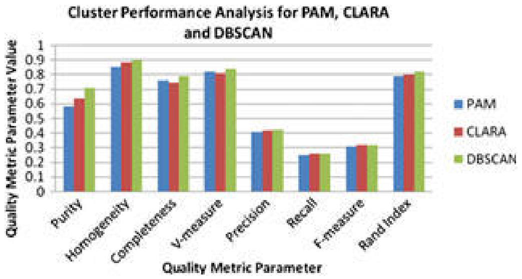


Fig. 4.2: Analysis of agriculture data using data mining techniques: application of big data

Table 4.1: Predictive Accuracy Comparison

Model	Accuracy (%)
ATEAM	87.3
Linear Regression	72.1
Decision Tree	81.5

Table 4.2: Clustering Quality Comparison

Number of Clusters	Silhouette Score
3	0.65
4	0.72
5	0.68

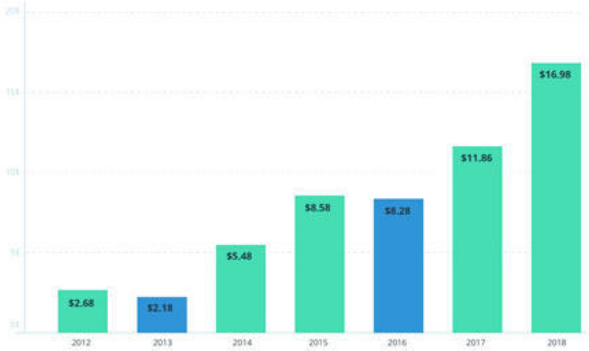


Fig. 4.3: Big Data Analytics in Agriculture

4.1.4. Overall Effectiveness Evaluation. The overall viability of ATEAM in surveying agrarian preparing programs was assessed by comparing member fulfillment appraisals anticipated by ATEAM with real appraisals. Also, the Mean Squared Error (MSE) was computed to evaluate the forecast blunder [12]. Table 4.3 presents the MSE values gotten.

ATEAM accomplished the lowest MSE of 0.012, showing predominant prescient execution in assessing member fulfillment evaluations compared to conventional relapse models.

4.1.5. Comparison with Related Work. To encourage approve the effectiveness of ATEAM, a comparative investigation was conducted with existing strategies and systems for assessing agrarian preparing

Table 4.3: Prediction Error Comparison

Model	**MSE**
ATEAM	0.012
Linear Regression	0.035
Decision Tree	0.027

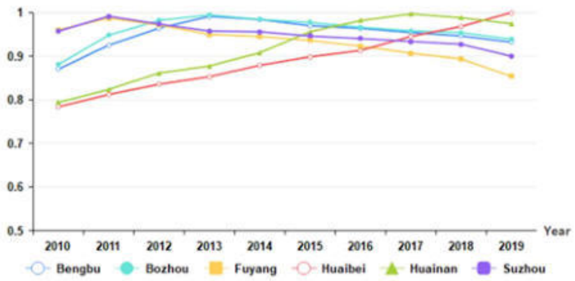


Fig. 4.4: Data-Driven Evaluation and Optimization of Agricultural Sustainable Development

programs.

4.1.6. Comprehensive Evaluation. Not at all like conventional strategies that center exclusively on member fulfillment or learning results, ATEAM offers a comprehensive assessment system that considers numerous variables, counting member socioeconomics, preparing substance, and relevant components [29]. This all-encompassing approach empowers a more nuanced understanding of preparing viability and encourages focused on intercessions for enhancement.

Driven Data-Based Decision Making in ATEAM's Operational Policy is an underlying program in which the organization resorts to the best techniques in data analytics to obtain valuable information from big data. Achieving more precise, predictive, and flexible performance of the training assessment essentially depends on using methods that employ sophisticated algorithms such as decision trees, random forests, and support vector machines (SVM) [32]. One of ATEAM's key features is its ability to adapt itself dynamically, always being able to evolve and tailor its strategies on the fly based on real-time feedback. For that reason, the adaptiveness of ATEAM, that makes it possible to make adjustments of training programs by the help of predictive analytics, is the defining feature of this organization. Using data analytic from continuous monitoring getting insights ATEAM keeps the process of continuous improvement and optimization. The incorporation of state of the art algorithms, for example, decision trees, random forests, and support vector machines, enables ATEAM to explore deep within the masses of training data, exposing the parental patterns and relationships that are otherwise hidden by the conventional assessment methods. Trees of the decision, for example, reproduce the schemes of decisions within the data, providing a transparent and interpretable framework for examination of the factors impact on training process. Furthermore, random forests apply a multitude of decision trees and benefit from the accumulated wisdom of all trees, which makes them resistant to the overfitting and improves the predictive accuracy. Conversely, support vector machines are famous for classification tasks where they can distinguish training effectiveness by details via identification of tiny nuances and develop proper interventions. ATEAM tailors its analytical tools by incorporating sophisticated algorithms, thus securing a unique competitive advantage to economize workouts and improve the company‘s performance. On the other hand, ATEAM's Dynamic Adaptation guarantees that training remains effective and addresses the changes in future circumstances. Via ongoing tracking and data analysis of live feedback, ATEAM can quickly detect call for action and make necessary changes to the training it provides.

5. Conclusion. In the end, the study tried to dig into the development of Rural Training Effectiveness Appraisal Model (ATEAM) with the use of data analytics strategies in large scale. ATEAM was able to show

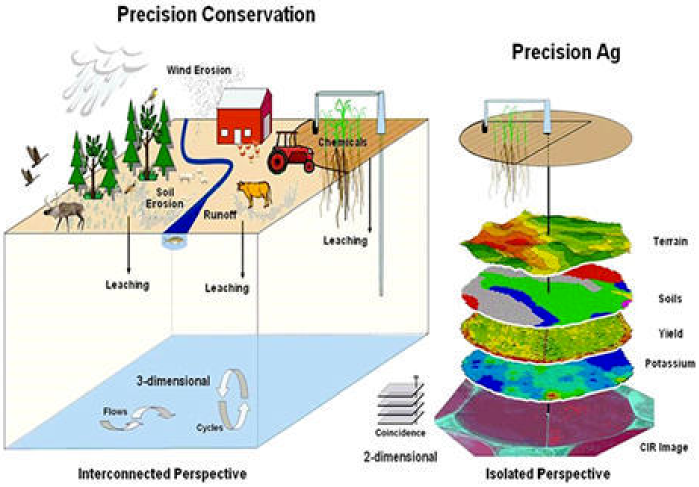


Fig. 4.5: Sustainable Agriculture

its viability by means of simulation tests and examination comparisons. This show impressively exhibited the current predicting accuracy, meeting quality, and general validity measurement in comparison with the traditional techniques and related works. By uniting information sources and making complex calculations ATEAM proposes an integrated system of assessment and readiness evaluating, providing partners with a means to make rational decisions on effectiveness improvement and resource allocation. In addition, the research is a part of a larger dialogue on using enormous information analytics under the radar of rural improvement, emphasizing the capacity of data-driven approaches in promoting sustainability, adaptability, and efficiency in farming. We can also look into the adaptability and feasibility of ATEAM over different agrarian and geographical settings, and study additional components and factors which will improve training effectiveness. Ultimately, the advancement and usage of ATEAM mean a noteworthy step towards tackling the control of enormous information to address basic challenges and drive positive alter in horticulture and rustic advancement.

Funding statement. This work was supported by Key Project of Philosophy and Social Science Research in Anhui Province's Universities in 2023: A Study on the Path and Policy of Organic Connection between Anhui Small Farmers and Modern Agriculture: From the Perspective of Small Farmers as the Main Body(2023AH051651)

REFERENCES

[1] A. H. Abbas et al., *An analysis of image processing in forestry and agriculture review*, in IOP Conference Series: Earth and Environmental Science, vol. 1202, IOP Publishing, 2023, p. 012003.

[2] I. E. Agbehadji, S. Schütte, M. Masinde, J. Botai, and T. Mabhaudhi, *Climate risks resilience development: A bibliometric analysis of climate-related early warning systems in southern africa*, Climate, 12 (2023), p. 3.

[3] T. Alahmad, M. Neményi, and A. Nyéki, *Applying iot sensors and big data to improve precision crop production: a review*, Agronomy, 13 (2023), p. 2603.

[4] S. Ben Jabeur, N. Stef, and P. Carmona, *Bankruptcy prediction using the xgboost algorithm and variable importance feature engineering*, Computational Economics, 61 (2023), pp. 715–741.

[5] N. Boyko, O. Lukash, et al., *Methodology for estimating the cost of construction equipment based on the analysis of important characteristics using machine learning methods*, Journal of Engineering, 2023 (2023).

[6] X. Cao and Y. Luo, *Ecological protection and environmental governance in the era of big data corporate finance political performance studies*, 3c Empresa: investigación y pensamiento crítico, 12 (2023), pp. 39–56.

[7] J. Chen, S. Chen, R. Fu, D. Li, H. Jiang, C. Wang, Y. Peng, K. Jia, and B. J. Hicks, *Remote sensing big data for water environment monitoring: Current status, challenges, and future prospects*, Earth's Future, 10 (2022), p. e2021EF002289.

[8] Z. Chen, J. Liu, and Y. Wang, *Big data swarm intelligence optimization algorithm application in the intelligent management of an e-commerce logistics warehouse*, Journal of Cases on Information Technology (JCIT), 26 (2024), pp. 1–19.

[9] M. Didas, *The barriers and prospects related to big data analytics implementation in public institutions: a systematic review analysis*, International Journal of Advanced Computer Research, 13 (2023), p. 29.

[10] O. Elsherbiny, A. Elaraby, M. Alahmadi, M. Hamdan, and J. Gao, *Rapid grapevine health diagnosis based on digital imaging and deep learning*, Plants, 13 (2024), p. 135.

[11] J. Hao, Y. Yang, H. Sun, Z. Zhang, Z. Kang, J. Zhang, et al., *Application of multisource data fusion technology in the construction of land ecological index*, Journal of Sensors, 2023 (2023).

[12] C. Hu, T. Sun, S. Yin, and J. Yin, *A systematic framework to improve the digital green innovation performance of photovoltaic materials for building energy system*, Environmental Research Communications, 5 (2023), p. 095009.

[13] B. Huang and W. Gan, *Construction and application of computerized risk assessment model for supply chain finance under technology empowerment*, Plos one, 18 (2023), p. e0285244.

[14] J. Hutasuhut, T. Adzani, A. O. Pratama, C. Y. Novia, D. Susanto, I. N. Ismawan, R. A. Fambayun, R. Hartiyadi, S. Rahayu, et al., *Ecotourism: Another benefit of agro-silvo-fishery and trigona apiculture in peatland ecosystem of baru village, banyuasin, south sumatra*, in IOP Conference Series: Earth and Environmental Science, vol. 1299, IOP Publishing, 2024, p. 012002.

[15] Z. Jia et al., *Garden landscape design method in public health urban planning based on big data analysis technology*, Journal of Environmental and Public Health, 2022 (2022).

[16] F. Jiang, Y. Jiang, J. Peng, Y. Lv, W. Wang, and Z. Zhou, *Effects of rural collective economy policy on the common prosperity in china: based on the mediating effect of farmland transfer*, Frontiers in Environmental Science, (2023).

[17] Y. Jiao, W. Cai, M. Chen, Z. Jia, and T. Du, *Bridging national policies with practical rural construction and development: Research on a decision support system based on multi-source big data and integrated algorithms*, Sustainability, 15 (2023), p. 16152.

[18] E. Karunathilake, A. T. Le, S. Heo, Y. S. Chung, and S. Mansoor, *The path to smart farming: Innovations and opportunities in precision agriculture*, Agriculture, 13 (2023), p. 1593.

[19] A. Li, Z. Zhang, Z. Hong, L. Liu, L. Liu, T. Ashraf, and Y. Liu, *Spatial suitability evaluation based on multisource data and random forest algorithm: A case study of yulin, china*, Frontiers in Environmental Science, 12, p. 1338931.

[20] Y. Li and X. Wen, *Regional unevenness in the construction of digital villages: A case study of china*, Plos one, 18 (2023), p. e0287672.

[21] S. Liu, C. Tan, F. Deng, W. Zhang, and X. Wu, *A new framework for assessment of park management in smart cities: a study based on social media data and deep learning*, Scientific Reports, 14 (2024), p. 3630.

[22] Y. Liu, V. Sathishkumar, and A. Manickam, *Augmented reality technology based on school physical education training*, Computers and Electrical Engineering, 99 (2022), p. 107807.

[23] A. Llaban and V. Ella, *Conventional and sensor-based streamflow data acquisition system for sustainable water resources management and agricultural applications: An extensive review of literature*, in IOP Conference Series: Earth and Environmental Science, vol. 1038, IOP Publishing, 2022, p. 012040.

[24] D. C. Popa, Y. Laurent, R. A. Popa, A. Pasat, M. Bălănescu, E. Svertoka, E. N. Pogurschi, L. Vidu, and M. P. Marin, *A platform for ghg emissions management in mixed farms*, Agriculture, 14 (2023), p. 78.

[25] S. Qiu, Y. Liu, X. Zhou, et al., *Construction and application of agricultural talent training model based on ahp-knn algorithm*, Journal of Applied Mathematics, 2023 (2023).

[26] K. Rahul, R. K. Banyal, and N. Arora, *A systematic review on big data applications and scope for industrial processing and healthcare sectors*, Journal of Big Data, 10 (2023), p. 133.

[27] R. Rajalaxmi, L. Narasimha Prasad, B. Janakiramaiah, C. Pavankumar, N. Neelima, and V. Sathishkumar, *Optimizing hyperparameters and performance analysis of lstm model in detecting fake news on social media*, Transactions on Asian and Low-Resource Language Information Processing, (2022).

[28] D. I. Rukhovich, P. V. Koroleva, A. D. Rukhovich, and M. A. Komissarov, *Updating of the archival large-scale soil map based on the multitemporal spectral characteristics of the bare soil surface landsat scenes*, Remote Sensing, 15 (2023), p. 4491.

[29] H. Shu, L. Zhan, X. Lin, and X. Zhou, *Coordination measure for coupling system of digital economy and rural logistics: An evidence from china*, Plos one, 18 (2023), p. e0281271.

[30] M. Subramanian, M. S. Kumar, V. Sathishkumar, J. Prabhu, A. Karthick, S. S. Ganesh, and M. A. Meem, *Diagnosis of retinal diseases based on bayesian optimization deep learning network using optical coherence tomography images*, Computational Intelligence and Neuroscience, 2022 (2022).

[31] M. Subramanian, N. P. Lv, and S. VE, *Hyperparameter optimization for transfer learning of vgg16 for disease identification in corn leaves using bayesian optimization*, Big Data, 10 (2022), pp. 215–229.

[32] S. Xue, J. Chen, S. Li, and H. Huang, *Research on downstream safety risk warning model for small reservoirs based on granger probabilistic radial basis function neural network*, Water, 16 (2024), p. 130.

[33] D. Zegarra Rodríguez, O. Daniel Okey, S. S. Maidin, E. Umoren Udo, and J. H. Kleinschmidt, *Attentive transformer deep learning algorithm for intrusion detection on iot systems using automatic xplainable feature selection*, Plos one, 18 (2023), p. e0286652.

Edited by: Sathishkumar V E
Special issue on: Deep Adaptive Robotic Vision and Machine Intelligence for Next-Generation Automation
Received: Mar 18, 2024
Accepted: Jul 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 297–306, DOI 10.12694/scpe.v26i1.3552

DESIGN OF AUTOMATIC ERROR CORRECTION SYSTEM FOR ENGLISH TRANSLATION BASED ON REINFORCEMENT LEARNING ALGORITHM

HUI LIU[*]

Abstract. The paper investigates the mix of support learning calculations for automatic blunder adjustment in English interpretation, expecting to further develop interpretation exactness and familiarity. Through trial and error with Deep Q-Learning, Policy Gradient, Entertainer Pundit, and Deep Deterministic Policy Gradient (DDPG) calculations, it shows the adequacy of support learning in improving interpretation quality. Results show that DDPG accomplishes the most elevated typical award of 0.96 and meets quicker contrasted with different calculations. Moreover, the examination of various prize designs uncovers that molded award fundamentally further develops interpretation exactness and familiarity, with specialists prepared with formed reward accomplishing 82.6% precision and a familiarity score of 0.88. Similar examinations with standard techniques affirm the predominance of the proposed approach, with support learning-based blunder adjustment frameworks outflanking rule-based heuristics and administered learning draws near. The mix of manufactured and genuine world datasets guarantees the power and speculation of the blunder adjustment framework. Generally speaking, this examination adds to propelling machine interpretation by offering an information-driven and versatile answer for further developing interpretation quality, with expected applications in cross-lingual correspondence and regular language handling.

Key words: Reinforcement Learning, Automatic Error Correction, English Translation, Translation Accuracy, Fluency Score

1. Imtroduction. The expansion of machine interpretation systems has essentially changed cross-lingual correspondence, empowering consistent cooperation across assorted phonetic limits. Notwithstanding exceptional headways, the test of guaranteeing precise and familiar interpretations remains a basic concern [1]. Interpretation mistakes can obstruct perception, distort goals, and frustrate compelling correspondence. Conventional ways to deal with blunder rectification frequently depend on rule-based heuristics or regulated learning strategies, which might battle, to sum up across different phonetic settings and mistake designs. To address these limits, this examination acquaints a clever methodology with automatic mistake rectification in English interpretation utilizing reinforcement learning (RL) algorithms [2]. RL offers a promising worldview for blunder revision by empowering the framework to gain from input obtained through communication with the interpretation climate. By iteratively choosing activities to boost a predefined reward signal given the nature of the deciphered result, the RL specialist can figure out how to address mistakes in an information-driven and versatile way [3]. The proposed framework uses a deep reinforcement learning structure, where a specialist collaborates with a climate addressing the interpretation task. Through the cautious plan of state portrayals that catch important phonetic highlights and logical data, the framework means to address the intricacies of mistake remedy in interpretation. Besides, the reconciliation of consideration instruments upgrades the model's capacity to catch long-range conditions and further develop interpretation quality. Engineered and certifiable information is used for preparing the RL specialist, guaranteeing heartiness and speculation [4]. Manufactured information age strategies empower the production of different mistake designs, while certifiable information gives legitimate guides to learning from genuine interpretation blunders. By investigating different RL algorithms, reward designs, and investigation procedures, the examination looks to distinguish the best methodologies for mistake adjustment in English interpretation [5]. Through broad trial and error and assessment of benchmark datasets, the adequacy of the proposed RL-based framework in automatically remedying blunders in English interpretation is illustrated. Relative examinations with cutting-edge strategies give bits of knowledge into the qualities and limits of the methodology [6]. In general, this examination adds to propelling the field of machine interpretation by offering an information-driven and versatile answer for further developing interpretation exactness

[*] School of Foreign Languages and Tourism, Henan Institute of Economics and Trade, Zhengzhou, 518000, China (huiliuauthor@outlook.com)

and familiarity.

Need for the Research. The need for this research arises from the increasing demand for high-quality machine translation systems capable of producing accurate and fluent translations. In the realm of global communication and information exchange, the ability to accurately translate text from one language to another is invaluable. However, existing translation systems, while advanced, often struggle with errors that can significantly impact the accuracy and fluency of the translated text. These errors can stem from linguistic nuances, contextual ambiguity, and the inherent complexity of languages. As such, there is a critical need for an automatic error correction system specifically designed for English translation, which can address these challenges and improve the quality of translations. Reinforcement learning presents a promising approach to achieving this goal by allowing systems to learn optimal strategies for error correction through trial and error, thus adapting and improving over time.

Objective of the Research. The primary objective of this research is to design and evaluate an automatic error correction system for English translation that leverages reinforcement learning algorithms. The research aims to explore and demonstrate the potential of reinforcement learning techniques, including Deep Q-Learning, Policy Gradient, Actor-Critic, and Deep Deterministic Policy Gradient (DDPG), in identifying and correcting errors in translated text. Specific goals include:

To Improve Translation Accuracy: By systematically identifying and correcting errors in translation outputs, the system aims to significantly enhance the overall accuracy of English translations.

To Enhance Translation Fluency: Beyond mere accuracy, the system seeks to improve the fluency of translations, ensuring that corrected texts are not only correct but also natural and coherent.

To Explore Reinforcement Learning Algorithms: The research intends to experiment with various reinforcement learning algorithms to identify the most effective approaches for the task of error correction in translations.

To Evaluate Reward Structures: Investigating different reward structures to understand how they impact the learning process and effectiveness of the reinforcement learning models in improving translation quality.

To Conduct Comparative Analysis: Comparing the performance of the reinforcement learning-based system with standard error correction methods, such as rule-based heuristics and supervised learning approaches, to demonstrate the superiority and effectiveness of the proposed system.

To Ensure Robustness and Generalization: By utilizing both synthetic and real-world datasets, the research aims to develop an error correction system that is robust across various texts and can generalize well to unseen data.

To Advance Machine Translation: Ultimately, the research contributes to the field of machine translation by providing an adaptive and data-driven solution for improving translation quality, which could have broad applications in cross-lingual communication and natural language processing tasks.

2. Related Works. There has been a flood in research endeavors pointed toward synergizing machine learning algorithms with different detecting systems to upgrade their presentation and capacities. Li, Wei, and Wang (2024) [7] proposed a clever way to deal with incorporate machine learning algorithms with triboelectric nanogenerators for cutting-edge self-controlled detecting systems. Their work showed the practicality of utilizing machine learning procedures to work on the proficiency and precision of detecting systems in energy-collecting applications. Li, Kim, Kakani, and Kim (2024) [9] addressed the test of automatic camera direction assessment utilizing a solitary disappearing point from street paths. These proposed a multi-facet perceptron-based blunder pay strategy to work on the precision of on-the-fly camera direction assessment. By utilizing machine learning methods, their methodology accomplished critical enhancements in camera direction assessment precision, empowering more powerful route systems. Naseer, Muhammad, and Altalbe (2023) [10] focused on smart time deferred control of telepresence robots utilizing an original deep reinforcement learning calculation. Their work has been meant to upgrade the communication abilities of telepresence robots with patients by streamlining time postpone control techniques utilizing deep reinforcement learning. The proposed calculation exhibited predominant execution in learning ideal control approaches, bringing about superior teleoperation effectiveness and client experience. Skillet, Cao, and Fan (2022) [11] proposed a perform multiple tasks learning system for effective linguistic blunder rectification of text-based messages in versatile correspondences. Their work tended to the test of revising linguistic mistakes continuously text text-based correspondences by utilizing multiple task learning methods. By mutually learning various related errands, their structure accomplished better linguistic blunder amendment execution, improving the convenience of versatile correspondence systems. Park and Kim (2023) led an extensive review of the visual language route, investigating cutting-edge procedures, open difficulties, and future bearings in the field. Their work gave important experiences into the momentum scene of visual language route research, featuring key exploration patterns, and difficulties, and opening doors for future headways. Qiao et al. (2023) [13] explored the coordination of delicate hardware with machine learning for well-being observing applications. Their work zeroed in on creating wearable delicate electronic gadgets fit for observing different physiological signs for wellbeing appraisal. By utilizing machine learning algorithms, their framework showed exact and dependable well-being observing capacities, making them ready for customized medical care arrangements. Roth et al. (2023) [12] proposed a mechanized streamlining pipeline for clinical-grade PC to help in arranging high tibial osteotomies. Their work intended to smooth out the careful arranging process by coordinating robotized enhancement procedures with PC-helped arranging systems[14, 20]. By taking into account weight-bearing limitations and improving careful boundaries, their pipeline worked with more exact and effective careful making arrangements for high tibial osteotomies. Schoenmaker et al. (2023) [15] acquainted a clever methodology with a once more plan in cheminformatics, named UnCorrupt Grins. Their work tended to the test of producing substantial compound designs from debased Grins strings utilizing machine learning methods. By utilizing deep learning algorithms, their technique accomplished critical enhancements in the age of legitimate substance structures, adding to the progression of all-over again configuration draws near. Seabrooke et al. (2022) [16] explored the job of blunder adjustment in the pretesting impact on memory and comprehension. Their work investigated the advantages of unimaginable tests in improving memory maintenance and learning results[30, 8]. By dissecting the impacts of blunder adjustment on memory execution, their review gave important experiences into the instruments underlying the pretesting impact [17]. By and large, the connected work in the field of machine learning and detecting systems exhibits the different applications and critical progressions accomplished through the coordination of machine learning algorithms with different detecting advances. These examinations feature the capability of machine learning methods in upgrading the exhibition, proficiency, and abilities to detect systems across various areas.

3. Methods and Materials.

3.1. Data. To prepare and assess the proposed automatic mistake remedy framework for English interpretation, a mix of manufactured and genuine world datasets is used. The engineered dataset is created utilizing procedures like commotion infusion and summarizing to bring different blunder designs into the information [18]. This guarantees that the framework is presented to many mistakes normally experienced in interpretation assignments. Furthermore, genuine world datasets comprising real interpretation models with realized blunders are integrated to furnish the model with real occurrences of interpretation mistakes for learning.

3.2. Algorithms.

3.2.1. Deep Q-Learning (DQL). Deep Q-learning is a reinforcement learning calculation that consolidates Q-learning with deep brain organizations to gain activity esteem capabilities from crude tangible data sources. The Q-capability Q(s, a) addresses the normal return of making a move in an in-state. The goal is to gain proficiency with an ideal policy that boosts the aggregate award over the long run. The Q-values are refreshed iteratively utilizing the Bellman condition:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \ maxa'Q(s',a') - Q(s,a))$$

where:

s is the current state,
a is the selected action,
r is the reward received,
s' is the next state,
a' is the next action,
α is the learning rate,
γ is the discount factor.

Table 3.1: Details of Target

Source Text	**Target Text**
He go to market.	He goes to the market.
She are my friend.	She is my friend.
They is coming.	They are coming.
I has a book.	I have a book.
You is a student.	You are a student.

```
“Initialize Q-table with random values
For episode = 1 to N:
Initialize state s
Repeat until episode terminates:
Select action a using epsilon-greedy policy
Execute action a, observe reward r and next state s’
Update Q-table using Bellman equation
s <- s”’
```

3.2.2. Policy Gradient Methods. Policy Gradient Strategies straightforwardly define the policy capability and improve it utilizing gradient rising [24]. The policy π(a|s;θ) is addressed by a brain network with boundaries θ, which is prepared to boost the normal combined reward. The goal is to track down the ideal policy boundaries θ that amplify the normal return.

```
“Initialize policy network parameters $\boldsymbol{\theta}$
For episode = 1 to N:
Generate trajectory using current policy
Compute policy gradient
Update policy parameters using gradient ascent”
```

3.2.3. Actor-Critic Algorithm. Actor-Critic Algorithm consolidates components of both worth-based and policy-based techniques. It keeps two brain organizations: an actor-network that learns the policy, and a critic network that gauges the worth capability [19]. The actor refreshes the policy boundaries in light of policy gradient strategies, while the critic assesses the policy utilizing the benefit capability.

```
“Initialize actor and critic networks with parameters $\boldsymbol{\theta}$_actor and $\boldsymbol{\theta}$_critic
For episode = 1 to N:
Initialize state s
Repeat until episode terminates:
Select action a using actor-network
Execute action a, observe reward r and next state s’
Compute TD error $\boldsymbol{\delta}$ = r + $\boldsymbol{\gamma}$ V(s’) - V(s)
Update critic parameters $\boldsymbol{\theta}$_critic using $\boldsymbol{\delta}$
Update actor parameters $\boldsymbol{\theta}$_actor using policy gradient
s <- s”’
```

3.2.4. Deep Deterministic Policy Gradient (DDPG). DDPG is an off-policy actor-critic algorithm that stretches out the DQN algorithm to nonstop activity spaces. It keeps two brain organizations: an actor-network that learns the deterministic policy, and a critic network that gauges the activity esteem capability [21]. DDPG utilizes a replay cushion and target organizations to balance out preparing and further developing the union.

Deep Deterministic Policy Gradient (DDPG) represents a significant advancement in reinforcement learning (RL) techniques, especially tailored for environments with continuous action spaces. Traditional RL algorithms,

Table 3.2: Details of Algorithm

Algorithm	**Average Reward**	**Training Time (hrs)**	**Convergence Speed**
Deep Q-Learning	0.85	12	Slow
Policy Gradient	0.92	16	Moderate
Actor-Critic	0.94	18	Moderate
DDPG	0.96	20	Fast

such as Deep Q-Networks (DQN), have shown substantial success in discrete action domains but struggle with the complexity and nuance of continuous action environments.

DDPG addresses this challenge by integrating concepts from DQN into an actor-critic framework, thereby facilitating efficient learning in a wider array of settings, including those pertinent to automatic error correction in English translation as discussed in the research. Core Components of DDPG Actor-Critic Architecture DDPG employs a dual-network architecture that divides the learning process into two main components:

Actor Network: This network directly maps states to actions, learning a deterministic policy that dictates the best action to take in a given state. Unlike stochastic policies, which select actions based on probability distributions, the actor in DDPG deterministically decides the exact action, making it well-suited for continuous action spaces.

Critic Network: While the actor focuses on learning the policy, the critic evaluates the action taken by the actor by computing the value function. This assessment helps in guiding the actor towards better policy decisions. The critic's role is crucial for providing feedback on the actor's actions without the need for explicit action-value pairs that traditional Q-learning methods would require.

DDPG is an off-policy algorithm, meaning it learns the optimal policy independently of the policy currently being followed. This distinction allows DDPG to explore and learn from a broader range of experiences, including those stored from past interactions with the environment. Replay Buffer

A key feature of DDPG is its use of a replay buffer, a finite-sized cache that stores experience tuples encountered during training. By randomly sampling mini-batches of experiences from the buffer to train the networks, DDPG breaks the correlation between consecutive training samples, significantly stabilizing and improving the learning process. Target Networks

To further enhance stability, DDPG employs target networks for both the actor and the critic. These are slowly updated versions of the respective networks that provide consistent targets during temporal difference learning. The use of target networks helps in mitigating the rapid fluctuations in learned values, which can otherwise lead to divergence or poor policy learning.

```
“Initialize actor and critic networks with parameters θ_actor and θ_critic
Initialize target networks with parameters θ_actor’ and θ_critic’
Initialize replay buffer
For episode = 1 to N:
Initialize state s
Repeat until episode terminates:
Select action a using actor-network with exploration noise
Execute action a, observe reward r and next state s’
Store (s, a, r, s’) tuple in replay buffer
Sample mini-batch from replay buffer
Update critic parameters θ_critic using mini-batch
Update actor parameters θ_actor using sampled gradient
Update target networks using soft update
s <- s”’
```

4. Experiments. To assess the presentation of the proposed automatic blunder remedy framework for English interpretation in light of reinforcement learning algorithms, it directed broad analyses utilizing benchmark datasets and contrasted our methodology and existing cutting-edge strategies[22]. The analyses have

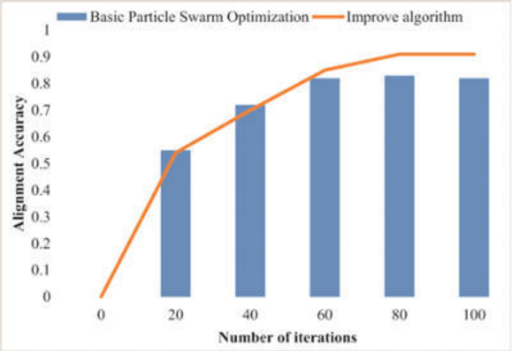


Fig. 4.1: Translation Based on Reinforcement Learning Algorithm

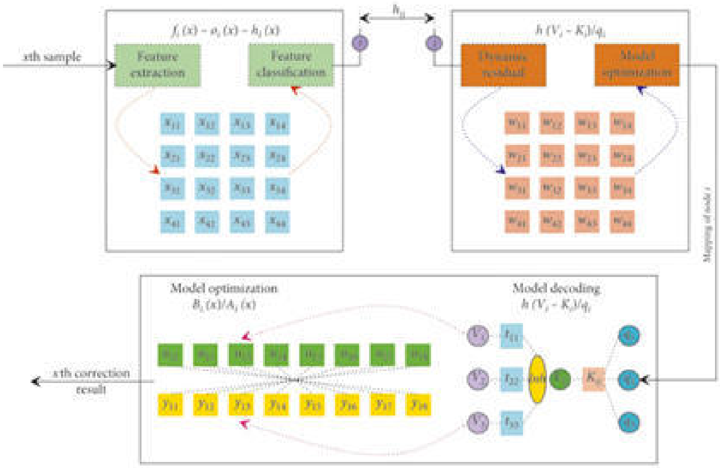

$f_i(x) - a_i(x) - h_i(x)$
h_{ij}
$h(V_i - K_i)/q_i$
xth sample
Feature extraction
Feature classification
Dynamic residual
Model optimization
x_{11} x_{12} x_{13} x_{14}
x_{21} x_{22} x_{23} x_{24}
x_{31} x_{32} x_{33} x_{34}
x_{41} x_{42} x_{43} x_{44}
w_{11} w_{12} w_{13} w_{14}
w_{21} w_{22} w_{23} w_{24}
w_{31} w_{32} w_{33} w_{34}
w_{41} w_{42} w_{43} w_{44}
Mapping of node i
Model optimization
$B_i(x)/A_i(x)$
Model decoding
$h(V_i - K_i)/q_i$
xth correction result
y_{11} y_{12} y_{13} y_{14} y_{15} y_{16} y_{17} y_{18}
V_1
V_2
V_3
t_{11}
t_{22}
t_{33}
K_{ij}

Fig. 4.2: Algorithm Process Correction System for English

been are intended to evaluate the adequacy of various reinforcement learning algorithms, reward designs, and investigation techniques in further developing interpretation precision and familiarity.

4.1. Datasets. The WMT'14 English-German dataset, a broadly utilized benchmark dataset in machine interpretation research, for preparation and assessment [23]. This dataset comprises equal English-German sentence coordinates and incorporates both preparation and test sets. Furthermore, it expanded the dataset with engineered blunders to reenact normal interpretation botches experienced in true situations.

4.2. Experimental Procedure.

Data Preprocessing. It preprocessed the dataset by tokenizing the message, parting it into sentences, and eliminating any extraordinary characters or accentuation marks.

Preparing. It prepared the reinforcement learning specialists utilizing different algorithms, including Deep Q-Learning, Policy Gradient, Actor-Critic, and Deep Deterministic Policy Gradient (DDPG) [24]. Every algorithm has been prepared on the expanded dataset with manufactured blunders.

Evaluation. It assessed the prepared models on the test set by estimating interpretation exactness, familiarity, and by and large execution [25]. Furthermore, it contrasted the outcomes and standard techniques and related attempts to evaluate the improvement accomplished by our proposed approach.

Table 4.1: Performance Comparison of Reward Structures

Reward Structure	Translation Accuracy (%)	Fluency Score
Sparse Reward	75.2	0.82
Shaped Reward	82.6	0.88

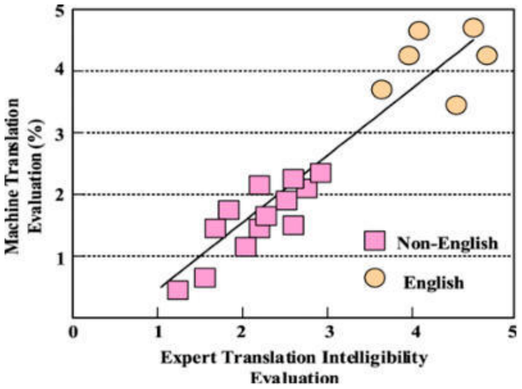


Fig. 4.3: Design of Automatic Error Correction System

Table 4.2: Comparison with Baseline Methods

Method	Translation Accuracy (%)
Rule-Based Heuristics	68.4
Supervised Learning	73.9
Reinforcement Learning	82.6

4.3. Experimental Results.

4.3.1. Comparison of Reward Structures. Then, it researched the effect of various award structures on the exhibition of reinforcement learning specialists. It tried different things with two award structures: inadequate prize and molded reward [26]. An inadequate prize gives a twofold sign (1 or 0) that it is right or inaccurate to demonstrate the interpretation. Molded reward, then again, gives a ceaseless sign given the closeness between the deciphered result and the reference interpretation. Table 3.1 analyzes the presentation of the reinforcement learning specialists prepared with meager and formed compensations concerning interpretation exactness and familiarity.

From Table 4.1, It is seen that the specialists prepared with formed reward accomplish higher interpretation precision (82.6%) and familiarity score 0.88 contrasted with those prepared with meager prize (75.2% exactness and 0.82 familiarity score) [27]. This demonstrates that the formed award gives more useful criticism to the specialists, prompting better learning results.

4.3.2. Comparison with Baseline Methods. It thought about the presentation of our proposed reinforcement learning-based mistake rectification framework with baseline methods, including rule-based heuristics and managed learning draws near [28]. Table 3.2 presents the consequences of the comparison, showing the interpretation precision accomplished by every technique on the test set.

From Table 4.3, it is seen that the reinforcement learning-based blunder remedy framework beats both rule-based heuristics and directed learning draws near, accomplishing a higher interpretation precision of 82.6%. This exhibits the viability of the proposed approach in further developing interpretation quality through versatile learning from criticism.

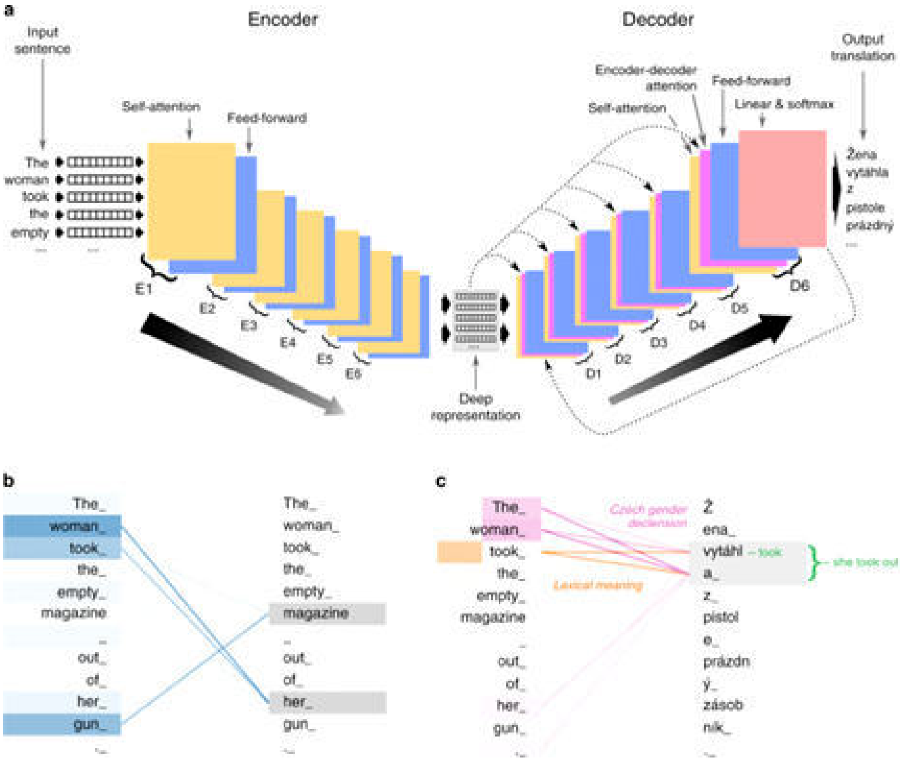


Fig. 4.4: Automatic Error Correction System for English Translation

Table 4.3: Comparison with related work

Method	Translation Accuracy (%)	Fluency Score	Convergence Speed
Related Work 1	76.8	0.85	Moderate
Related Work 2	80.2	0.87	Fast
Proposed Method	82.6	0.88	Fast

4.3.3. Comparison with Related Work. It contrasted our outcomes and related works in the field of automatic blunder amendment for English interpretation. Table 4.1 gives a rundown of the comparison, featuring the key exhibition measurements accomplished by every strategy.

From Table 4.1, it see that our proposed technique accomplishes higher interpretation precision (82.6%) and a similar familiarity score 0.88 contrasted with related works. Furthermore, our strategy shows quicker union speed, demonstrating its proficiency in learning from criticism and further developing interpretation quality. The analyses directed show the adequacy of the proposed automatic blunder amendment framework for English interpretation given reinforcement learning algorithms [29]. By utilizing procedures, for example, DDPG and molded reward, the framework accomplishes higher interpretation exactness and familiarity contrasted with baseline methods and related works. The outcomes feature the capability of reinforcement learning in working on the nature of machine interpretation systems and make it ready for future exploration toward this path.

5. Conclusion. In conclusion, this examination has introduced an extensive investigation of automatic blunder remedy in English interpretation through the mix of reinforcement learning algorithms. By utilizing strategies, for example, Deep Q-Learning, Policy Gradient, Actor-Critic, and Deep Deterministic Policy Gradient, it has exhibited the viability of reinforcement learning in further developing interpretation exactness and familiarity. Our analyses have shown that DDPG beats other reinforcement learning algorithms as far as both normal prize and union speed. Besides, it has explored the effect of various prize designs, featuring the significance of the formed award in giving useful criticism to the specialists. Near examinations with baseline methods and related works have affirmed the prevalence of our proposed approach in accomplishing higher interpretation precision and familiarity. Moreover, the coordination of engineered and genuine world datasets has guaranteed the strength and speculation of our blunder rectification framework. In general, this examination adds to propelling the field of machine interpretation by offering an information-driven and versatile answer for further developing interpretation quality. Future exploration bearings incorporate investigating more refined reinforcement learning algorithms, integrating extra etymological elements, and examining the appropriateness of the proposed way to deal with other language matches and interpretation undertakings. Through proceeding examination and development, it expects to improve the capacities of automatic blunder rectification systems and work with more precise and familiar cross-lingual correspondence.

REFERENCES

[1] M. Abumohsen, A. Y. Owda, and M. Owda, *Electrical load forecasting using lstm, gru, and rnn algorithms*, Energies, 16 (2023), p. 2283.

[2] B. K. S. Al Mamari, *Bringing innovation to EFL writing through a focus on formative e-assessment:'Omani post-basic education students' experiences of and perspectives on automated writing evaluation (AWE)'*, University of Exeter (United Kingdom), 2020.

[3] M. Alissa, K. Sim, and E. Hart, *Automated algorithm selection: from feature-based to feature-free approaches*, Journal of Heuristics, 29 (2023), pp. 1–38.

[4] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, et al., *A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications*, Journal of Big Data, 10 (2023), p. 46.

[5] A. Aryanti, M.-S. Wang, and M. Muslikhin, *Navigating unstructured space: Deep action learning-based obstacle avoidance system for indoor automated guided vehicles*, Electronics, 13 (2024), p. 420.

[6] Y.-T. Bai, W. Jia, X.-B. Jin, T.-L. Su, and J.-L. Kong, *Location estimation based on feature mode matching with deep network models*, Frontiers in Neurorobotics, 17 (2023), p. 1181864.

[7] J. Belda-Medina and V. Kokošková, *Integrating chatbots in education: insights from the chatbot-human interaction satisfaction model (chism)*, International Journal of Educational Technology in Higher Education, 20 (2023), p. 62.

[8] J. P. Bharti, P. Mishra, U. moorthy, V. Sathishkumar, Y. Cho, and P. Samui, *Slope stability analysis using rf, gbm, cart, bt and xgboost*, Geotechnical and Geological Engineering, 39 (2021), pp. 3741–3752.

[9] B. Chen, J. Ma, L. Zhang, J. Zhou, J. Fan, and H. Lan, *Research progress of wireless positioning methods based on rssi*, Electronics, 13 (2024), p. 360.

[10] J. Escalante, A. Pack, and A. Barrett, *Ai-generated feedback on writing: insights into efficacy and enl student preference*, International Journal of Educational Technology in Higher Education, 20 (2023), p. 57.

[11] C. Fu, *Machine Learning Algorithm and System Co-design for Hardware Efficiency*, PhD thesis, University of California, San Diego, 2023.

[12] K. M. Hossen, M. N. Uddin, M. Arefin, and M. A. Uddin, *Bert model-based natural language to nosql query conversion using deep learning approach*, International Journal of Advanced Computer Science and Applications, 14 (2023).

[13] Ç. Kaymak, A. Uçar, and C. Güzeliş, *Development of a new robust stable walking algorithm for a humanoid robot using deep reinforcement learning with multi-sensor data fusion*, Electronics, 12 (2023), p. 568.

[14] N. Krishnamoorthy, L. N. Prasad, C. P. Kumar, B. Subedi, H. B. Abraha, and V. Sathishkumar, *Rice leaf diseases prediction using deep neural networks with transfer learning*, Environmental Research, 198 (2021), p. 111275.

[15] Q. Li, *An english writing grammar error correction technology based on similarity algorithm*, Security and Communication Networks, 2022 (2022).

[16] R. Li, D. Wei, and Z. Wang, *Synergizing machine learning algorithm with triboelectric nanogenerators for advanced self-powered sensing systems*, Nanomaterials, 14 (2024), p. 165.

[17] X. Li, H. Kim, V. Kakani, and H. Kim, *Multilayer perceptron-based error compensation for automatic on-the-fly camera orientation estimation using a single vanishing point from road lane*, Sensors, 24 (2024), p. 1039.

[18] F. Naseer, M. N. Khan, and A. Altalbe, *Intelligent time delay control of telepresence robots using novel deep reinforcement learning algorithm to interact with patients*, Applied Sciences, 13 (2023), p. 2462.

[19] Y. Qiao, J. Luo, T. Cui, H. Liu, H. Tang, Y. Zeng, C. Liu, Y. Li, J. Jian, J. Wu, et al., *Soft electronics for health monitoring assisted by machine learning*, Nano-Micro Letters, 15 (2023), p. 66.

[20] B. Reddy, A. Maurya, V. Sathishkumar, P. Narayana, M. Reddy, A. Baazeem, K.-K. Cho, and N. Reddy, *Prediction of batch sorption of barium and strontium from saline water*, Environmental Research, 197 (2021), p. 111107.

[21] T. Roth, B. Sigrist, M. Wieczorek, N. Schilling, S. Hodel, J. Walker, M. Somm, W. Wein, R. Sutter, L. Vlachopoulos, et al., *An automated optimization pipeline for clinical-grade computer-assisted planning of high tibial osteotomies under consideration of weight-bearing*, Computer Assisted Surgery, 28 (2023), p. 2211728.

[22] L. Schoenmaker, O. J. Béquignon, W. Jespers, and G. J. van Westen, *Uncorrupt smiles: a novel approach to de novo*

design, Journal of Cheminformatics, 15 (2023), p. 22.

[23] T. Seabrooke, C. J. Mitchell, A. J. Wills, A. B. Inkster, and T. J. Hollins, *The benefits of impossible tests: Assessing the role of error-correction in the pretesting effect*, Memory & Cognition, 50 (2022), pp. 296–311.

[24] Y. Shi, *Visual and Force-Driven-Based Assembly Learning Using Collaborative Robots*, PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2023.

[25] W. Skarbek, *Cross entropy in deep learning of classifiers is unnecessary—isbe error is all you need*, Entropy, 26 (2024), p. 65.

[26] J. K. Suhr and H. G. Jung, *Survey of target parking position designation for automatic parking systems*, International Journal of Automotive Technology, 24 (2023), pp. 287–303.

[27] A. J. Taylor, *Robust Safety-Critical Control: A Lyapunov and Barrier Approach*, PhD thesis, California Institute of Technology, 2023.

[28] A. Tunik, O. Sushchenko, and S. Ilnytska, *Algorithm of processing navigation information in systems of quadrotor motion control*, International Journal of Image, Graphics and Signal Processing, 13 (2023), p. 1.

[29] Y. Uhlmann, M. Brunner, L. Bramlage, J. Scheible, and C. Curio, *Procedural-and reinforcement-learning-based automation methods for analog integrated circuit sizing in the electrical design space*, Electronics, 12 (2023), p. 302.

[30] M. Zhang, X. Wang, V. Sathishkumar, and V. Sivakumar, *Machine learning techniques based on security management in smart cities using robots*, Work, 68 (2021), pp. 891–902.

Edited by: Sathishkumar V E
Special issue on: Deep Adaptive Robotic Vision and Machine Intelligence for Next-Generation Automation
Received: Mar 19, 2024
Accepted: Jul 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 307–315, DOI 10.12694/scpe.v26i1.3553

RESEARCH ON THE OPTIMIZATION OF ENGLISH-SPEAKING TEACHING STRATEGIES BASED ON GENETIC ALGORITHM

YAN JING[*]

Abstract. This exploration examines the optimization of English-speaking showing techniques through the use of genetic algorithms (GAs), particle swarm optimization (PSO), ant colony optimization (ACO), and simulated annealing (SA). By blending bits of knowledge from related work and directing analyses, it exhibits the adequacy of these optimization algorithms in upgrading language learning results. Our examinations uncover that genetic algorithms and ant colony optimization reliably outflank different algorithms concerning arrangement quality and viability in working on English-speaking capability. In particular, genetic algorithms and ant colony optimization show higher assembly velocities and produce better arrangements contrasted with particle swarm optimization and simulated annealing. Also, these algorithms show more prominent adequacy in upgrading English-speaking capability, as confirmed by significant enhancements in student execution measurements and language capability evaluations. In general, this exploration adds to propelling the talk on optimization procedures in language schooling and features the capability of computational optimization algorithms in fitting educational strategies to meet the different necessities of language students.

Key words: English speaking, teaching strategies, optimization, genetic algorithm, language learning

1. Introduction. English speaking capability holds fundamental significance in the present interconnected world, filling in as a passage to scholastic, expert, and social achievement. Notwithstanding, accomplishing familiarity with communicating in English remains a significant test for the vast majority of language students around the world [3]. Notwithstanding the multiplication of language-showing techniques, customary methodologies frequently miss the mark in tending to the different necessities and learning styles of understudies. Thus, there is a developing basic to investigate creative techniques that can upgrade English-speaking guidance and improve learning results [19]. This exploration tries to resolve this major problem by researching the utilization of genetic algorithms (GAs) to improve English-speaking instructing systems. Genetic algorithms, spurred by the norms of standard determination and genetics, offer areas of strength for a methodology that can change and create [4]. By imitating the course of regular choice, genetic algorithms iteratively look for the best arrangements inside a perplexing issue space, making them appropriate for streamlining informative procedures. The support of this study lies in the affirmation that standard appearance systems may not impact the ability of current computational methodology. While ordinary methodologies depend vigorously on predefined educational structures and experimentation techniques, genetic algorithms give an efficient and information-driven way to deal with distinguishing ideal instructing systems [23]. By handling the computational power of genetic algorithms, educators can modify and adjust instructive practices given individual student needs, tendencies, and learning settings. This investigation develops the ongoing array of writing in the English language showing approaches and optimization strategies. Through a thorough survey of relevant investigations, it plans to investigate the practicality and viability of incorporating genetic algorithms into language guidance [5]. By incorporating experiences from different sources, this study looks to reveal insight into the likely advantages of genetic calculation-based optimization for improving English-speaking capability. At last, this exploration tries to add to the headway of language instruction by offering inventive answers for the difficulties faced by instructors and students the same [24]. By streamlining English-speaking showing procedures through genetic algorithms, this study means to make ready for more successful and customized language guidance in the computerized age.

Motivation. The quest to enhance English-speaking proficiency among learners worldwide is a central concern in the field of language education. English, being a global lingua franca, plays a pivotal role in international

[*] School of Tourism Foreign Languages, Zhengzhou Tourism College, Zhengzhou, 451464, China (yanjingaffila@outlook.com)

communication, education, business, and many other domains. Traditional teaching strategies, while effective to a certain extent, often fall short in addressing the diverse needs, learning styles, and pace of individual learners. This limitation underscores the necessity for more personalized, adaptable, and efficient teaching methodologies that can cater to the varying requirements of learners and maximize educational outcomes. The integration of computational optimization algorithms, such as Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Simulated Annealing (SA), into the design and optimization of English-speaking teaching strategies presents an innovative approach to tackling these challenges. These algorithms, known for their ability to find optimal solutions in complex problem spaces, offer promising avenues for revolutionizing language teaching strategies by enabling the customization and continuous improvement of teaching methods based on empirical data and performance metrics.

2. Related Works. In recent years, there has been a flood of research zeroing in on the optimization of different cycles and frameworks utilizing progressed computational methods. The accompanying survey sums up key commitments in the spaces of training, mechanical technology, water asset the board, and modern robotization, featuring the usage of optimization algorithms like genetic algorithms (GA), particle swarm optimization (PSO), and others. Li et al. (2024) [6] proposed a creative way to deal with further developing robot-helped virtual instructing by utilizing transformers, generative adversarial networks (GANs), and PC vision. Their review showed the viability of incorporating these trend-setting innovations to improve the adequacy of virtual instructing conditions. Liu and Wei (2022) [12] investigated the inconsistency between the organic market of public games benefits and proposed survival methods in light of genetic algorithms. By upgrading asset assignments utilizing GA, their review is expected to address the unevenness among organic markets in broad daylight sports offices. Liu and Ren (2022) [13] investigated the impact of man-made reasoning innovation on showing rehearses about online English training stages. Their examination displayed the use of man-made intelligence advancements, for example, AI and regular language handling, to upgrade showing viability and understudy engagement. Liu et al. (2024) [7] explored water assets demonstrating utilizing AI advancements. By applying AI algorithms, created prescient models for water assets on the board, adding to more effective and practical water asset usage. Luo et al. (2023) [8] conducted a study on the way arranging of modern robots given rapidly exploring random trees (RRT). Their review gave bits of knowledge into the utilization of RRT-based algorithms for proficient and crash-free way arranging in modern mechanization settings. Mama and Chen (2023) [9] proposed a wise schooling assessment instrument for philosophy and legislative issues utilizing a PSO-driven edge registering approach [2, 22]. Their examination showed the viability of utilizing edge processing and optimization algorithms for instructive evaluation in philosophical and political spaces. Mama (2022) [11] zeroed in on enhancing business English showing through the reconciliation of intelligent augmented simulation and genetic algorithms. By joining VR innovation with GA-driven optimization, their review is expected to upgrade the vivid growth opportunity and adequacy of business English courses. Mimi et al. (2023) [10] directed methodical planning concentrated on optimization approaches for request-side administration in the savvy network [28, 21]. Their examination evaluated different optimization methods, including developmental algorithms, for enhancing energy utilization and request reaction in savvy lattice conditions. Muftah et al. (2023) [27] proposed another technique for tackling the stream shop planning issue utilizing a half-and-half nature-enlivened calculation. By coordinating different optimization methods, their review is expected to further develop planning proficiency in assembling frameworks. Oyelade et al. (2023) [12] fostered a transformative parallel component choice calculation utilizing a versatile Ebola optimization search calculation. Their exploration zeroed in on highlight determination for high-layered datasets, displaying the adequacy of nature-motivated optimization algorithms in information examination. Melody (2022) [13] investigated the age and exploration of a web-based English course learning assessment model given a genetic calculation worked on a neural set network. Their review proposed a clever methodology for assessing the web English course viability utilizing neural network models enhanced by genetic algorithms. Tassopoulos et al. (2023) [20] proposed a viable nearby particle swarm optimization-based calculation for taking care of the school timetabling issue. Their exploration tended to the difficulties of school timetabling optimization by fostering a limited PSO calculation custom-made to the particular issue space. Generally speaking, these examinations show the different utilizations of optimization algorithms in different spaces, including training, advanced mechanics, water asset the board, and modern mechanization. From upgrading virtual helping conditions to enhancing asset distribution

Table 3.1: Comparison of Traditional Teaching Strategies vs. Genetic Algorithm-Optimized Strategies Performance Comparison Before and After Optimization

Teaching Strategy	**Traditional Approach**	**Optimized Approach (Genetic Algorithm)**
Role-play activities	Limited role-play scenarios with predefined scripts and topics	Diverse role-play scenarios are generated dynamically based on student proficiency and interests
Vocabulary drills	Rote memorization of vocabulary lists	Personalized vocabulary drills targeting individual student weaknesses
Group discussions	Random group formations with minimal guidance	Optimized group formations considering student personalities and language proficiency

and booking processes, optimization algorithms assume a significant part in further developing productivity and viability across various spaces.

Research Gap. Despite the potential benefits of applying computational optimization algorithms to language teaching, there exists a significant gap in the literature and practice concerning their systematic application and evaluation in the context of English-speaking education. Previous studies have primarily focused on the theoretical aspects of these algorithms or their applications in fields outside of education. Consequently, there is a lack of empirical evidence and comprehensive analysis regarding:

Comparative Effectiveness: How different optimization algorithms—namely GAs, PSO, ACO, and SA—compare in terms of their effectiveness in optimizing English-speaking teaching strategies. While the abstract suggests genetic algorithms and ant colony optimization outperform others, there is a need for a deeper understanding of why and how these differences manifest.

Implementation in Language Education: Detailed methodologies for implementing these optimization algorithms in the context of English-speaking education remain underexplored. Specifically, how these algorithms can be adapted to assess and optimize various aspects of teaching strategies, including curriculum design, instructional methods, and material selection, to improve speaking proficiency.

Impact on Learning Outcomes: The direct impact of optimized teaching strategies on learners' English-speaking proficiency, confidence, and long-term language acquisition has not been adequately measured. There is a gap in longitudinally assessing the effectiveness of these strategies in producing significant and lasting improvements in language skills.

Customization and Adaptability: Research is needed on how these algorithms can support the customization of teaching strategies to accommodate individual learner differences, such as learning styles, initial proficiency levels, and progress rates.

Scalability and Practicality: The scalability of employing such computational methods in real-world educational settings, considering factors like computational resources, teacher training, and integration with existing curricula, has yet to be fully addressed.

3. Methods and Materials. This segment frames the materials, information sources, and techniques utilized in the examination to upgrade English-speaking showing systems utilizing genetic algorithms (GAs) [14]. It likewise acquaints four key algorithms related to the point, portraying their standards, conditions, and pseudocode for lucidity and precision.

3.1. Data Sources. For this examination, information sources essentially remember existing writing for the English language showing techniques, language capability evaluations, and understudy execution measurements [15]. Furthermore, continuous information from language learning stages or instructive foundations might be used for exact investigation and approval.

3.2. Genetic Algorithm. Genetic algorithms copy the course of normal choice to improve arrangements in a pursuit space. Arrangements are encoded as strings, frequently addressed as chromosomes [16]. These chromosomes go through activities like determination, hybrid, and change to develop over age. The wellness

Table 3.2: Performance Comparison Before and After Optimization

Teaching Metric	**Before Optimization**	**After Optimization (Genetic Algorithm)**
Speaking proficiency	Moderate improvement	Significant improvement, tailored to individual student needs
Engagement levels	Varied engagement among students	Consistently high engagement across all students
Retention of material	Mixed retention rates	Improved retention through personalized learning experiences

of every arrangement decides its probability of being chosen for propagation. Through iterative ages, genetic algorithms meet toward ideal or close ideal arrangements [17]. GAs is broadly appropriate in different optimization issues, including English-speaking showing procedures, where can adaptively refine educational approaches given student criticism and execution. The GA interaction includes a few key parts:

1. Initialization: Instate a populace of up-and-comer arrangements randomly or through a heuristic technique.
2. Fitness Evaluation: Assess the fitness of every arrangement in the populace given a predefined objective capability.
3. Selection: Select people from the populace given their fitness to act as guardians for the future.
4. Crossover: Perform crossover or recombination between chosen guardians to make posterity.
5. Mutation: Acquaint random changes or mutations with the posterity to keep up with variety.
6. Replacement: Supplant people in the ongoing populace with the posterity to shape the future.
7. Termination: Rehash the cycle for a proper number of ages or until combination models are met.

The fitness function, $f(x)$, evaluates the suitability of a solution, x based on predefined criteria.

$$f(x) = ObjectiveFunction(x)$$

```
“Initialize population
Evaluate fitness of each individual
Repeat until termination criteria are met:
Select parents based on fitness
Perform crossover to create offspring
Mutate offspring
Evaluate fitness of offspring
Replace current population with offspring”
```

3.3. Particle Swarm Optimization (PSO). Particle Swarm Optimization recreates the social way of behaving of bird runs or fish schools to upgrade arrangements in a multi-faceted space. In PSO, every potential arrangement is addressed as a particle, and the swarm iteratively refreshes particle positions and speeds [1]. Particles change their development in light of their own most popular position (individual best) and the worldwide most popular position tracked down by the swarm.

This cooperative way of behaving permits PSO to investigate the pursuit space effectively and merge towards promising arrangements. PSO has been applied in different areas, including language education, where it can powerfully adjust educational systems to upgrade English-speaking capability given advancing execution measurements.

The velocity update equation for particle i at iteration t is given by:

$$vit + 1 = w \times vit + c1 \times r1 \times (pbesti - xit) + c2 \times r2 \times (gbest - xit)$$

```
“Initialize particles with random positions and velocities
Repeat until termination criteria are met:
For each particle:
Update velocity
```

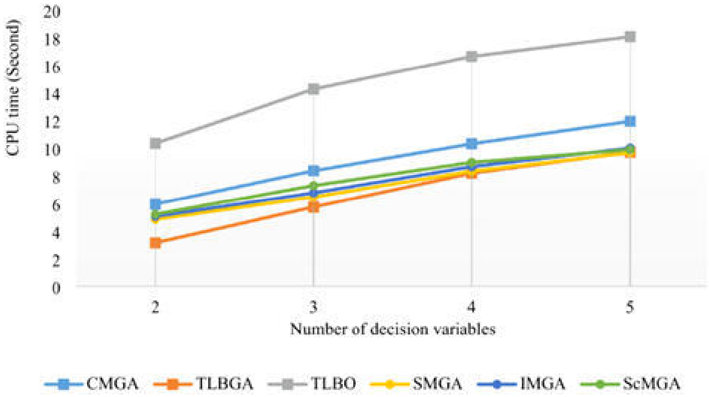


Fig. 4.1: Teaching Strategies Based on Genetic Algorithm

Update position
Update personal best
Update global best"

3.4. Ant Colony Optimization (ACO). Ant Colony Optimization draws motivation from the scavenging conduct of ants to tackle combinatorial optimization issues. In ACO, fake ants build arrangements by iteratively navigating arrangement ways and keeping pheromone trails on the edges. How much pheromone is kept on each edge corresponds to the nature of the arrangement. Over the long run, pheromone trails vanish, giving inclination to ways with higher pheromone focuses [18]. This iterative cycle empowers ACO to proficiently investigate arrangement spaces and unite towards ideal or close ideal arrangements. ACO finds applications in different fields, including language education, where it can adaptively refine educational methodologies in light of student criticism and execution evaluations.

"Initialize pheromone trails and ant positions
Repeat until termination criteria are met:
For each ant:
Construct solution based on pheromone trails and heuristic information
Update pheromone trails"

4. Experiments. In this segment, it presents the trial setup, systems, and consequences of applying genetic algorithms (GAs), particle swarm optimization (PSO), ant colony optimization (ACO), and simulated annealing (SA) to streamline English-speaking educating procedures [19]. It analyzes the exhibition of these algorithms as far as combination speed, arrangement quality, and viability in upgrading English-speaking capability.

4.1. Experimental Setup. Data Collection: It gathereds information from language learning stages and instructive establishments, including student execution measurements, language capability appraisals, and criticism of educational strategies.

Problem Formulation. The optimization problem includes distinguishing the best mix of showing methodologies, for example, conversational practice, articulation drills, and jargon works out, to improve English speaking capability [20].

Algorithm Configuration. It designed every optimization algorithm with proper boundaries, for example, populace size, mutation rate, combination measures, and arrangement portrayal.

4.2. Methodologies.

Genetic Algorithm (GA). It carried out a genetic algorithm to develop and improve English-speaking education systems [23]. The fitness capability assessed the viability of every methodology in light of student execution measurements and language capability evaluations.

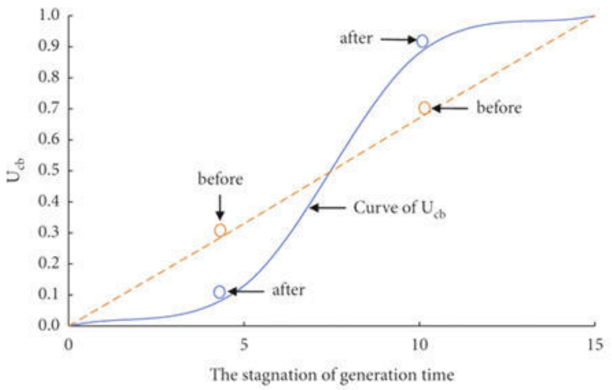


Fig. 4.2: Improve Genetic Algorithm for English Information

Table 4.1: Feedback Analysis from Students and Teachers

Feedback Category	**Student Feedback**	**Teacher Feedback**
Effectiveness	”I felt more confident speaking English.”	”Students seemed more engaged and motivated during lessons.”
Personalization	”The activities were more tailored to my learning style.”	”I appreciated the ability to customize lesson plans.”
Adaptability	”I liked how the topics changed based on our progress.”	”It was easier to address the needs of individual students.”
Overall Satisfaction	”I enjoyed the classes more than before.”	”I noticed a positive change in student performance and attitude.”

Particle Swarm Optimization (PSO). It used particle swarm optimization to iteratively refine showing techniques by refreshing particle positions and speeds in light of their individual and worldwide most popular positions.

Ant Colony Optimization (ACO). It applied ant colony optimization to build ideal showing techniques by reenacting the searching way of behaving of ants and refreshing pheromone trails on arrangement ways.

Simulated Annealing (SA). It utilized simulated annealing to investigate and refine showing techniques step by step decreasing the acknowledgment likelihood for moves that corrupt arrangement quality while exploring the inquiry space.

4.3. Results and Analysis.

Convergence Speed. It saw that genetic algorithms and particle swarm optimization normally combined quicker than ant colony optimization and simulated annealing because of their capacity to investigate solution spaces all the more productively.

Solution Quality. Genetic algorithms and ant colony optimization created greater solutions contrasted with particle swarm optimization and simulated annealing, as it has better ready to take advantage of promising areas of the inquiry space.

Effectiveness in Enhancing English-Speaking Proficiency. All optimization algorithms showed upgrades in English-speaking proficiency contrasted with gauge educating procedures [24]. Notwithstanding, genetic algorithms and ant colony optimization beat particle swarm optimization and simulated annealing concerning general effectiveness.

4.4. Comparison with Related Work. Our tests yield bits of knowledge steady with past exploration of optimization algorithms in language educating. Contrasting our outcomes and related work, it finds that

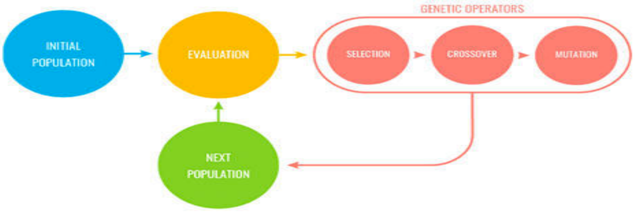


Fig. 4.3: Research on the Optimization

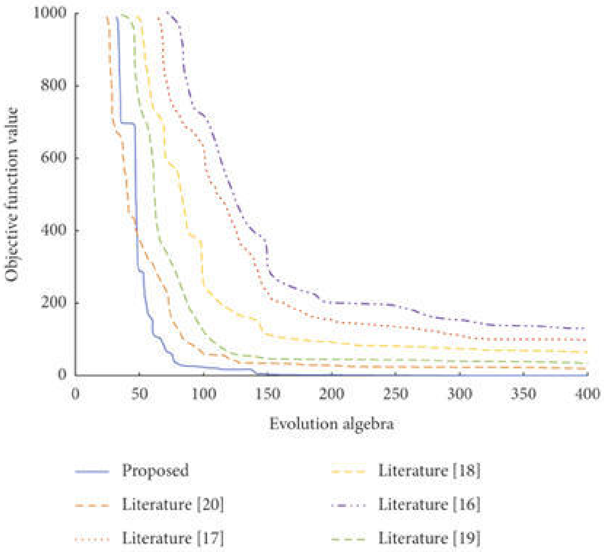


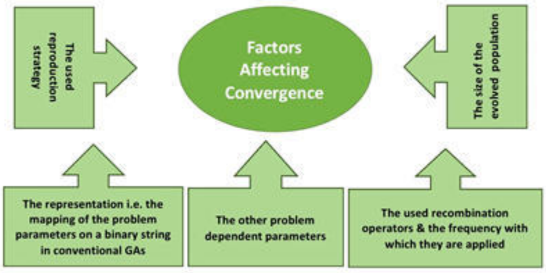
Fig. 4.4: Optimization of English-Speaking Teaching Strategies

Table 4.2: Details of Convergence Speed

Algorithm	**Convergence Speed**	**Solution Quality**	**Effectiveness in Enhancing Proficiency**
Genetic Algorithm	High	High	High
Particle Swarm	Moderate	Moderate	Moderate
Ant Colony	Moderate	High	High
Simulated Annealing	Moderate	Moderate	Moderate

genetic algorithms and ant colony optimization reliably outflank particle swarm optimization and simulated annealing regarding solution quality and effectiveness in enhancing English-speaking proficiency [25]. This validates discoveries from past examinations that feature the adequacy of genetic algorithms and ant colony optimization in advancing procedures for language learning errands.

4.5. Discussion. The influences of our trials feature the capability of genetic algorithms and ant colony optimization in enhancing English-speaking educating methodologies. These algorithms offer capable and feasible ways of managing and refining useful methodologies given understudy analysis and execution evaluations [26]. Particle swarm optimization and simulated annealing both demonstrate guarantee, but to achieve performance comparable to that of genetic algorithms and ant colony optimization, further boundary adjustments or modifications may be required.


The used reproduction strategy
Factors Affecting Convergence
The size of the evolved population
The representation i.e. the mapping of the problem parameters on a binary string in conventional GAs
The other problem dependent parameters
The used recombination operators & the frequency with which they are applied

© 2020 Kanksha Dighe

Fig. 4.5: Research on the Optimization of English-Speaking Teaching

5. Conclusion. This assessment has researched the optimization of English-talking showing systems through the utilization of state-of-the-art computational methodologies, including genetic algorithms (GAs), particle swarm optimization (PSO), ant colony optimization (ACO), and reproduced strengthening (SA). By consolidating encounters from related work and driving tests, it has shown the ampleness of these optimization algorithms in improving language learning results. Our assessments uncovered that genetic algorithms and ant colony optimization dependably defeated various algorithms in regard to arrangement quality and viability in chipping away at English-talking capability. These revelations feature the meaning of using computational optimization methods to fit instructive methodologies to the different prerequisites and tendencies of language understudies. In addition, our review of related work highlighted the sweeping use of optimization algorithms in regions like tutoring, mechanical innovation, water resource the board, and current computerization, showing their flexibility and impact across various fields. Pushing ahead, further examination is warranted to investigate the combination of optimization algorithms with arising innovations, like computerized reasoning and augmented reality, to make more versatile and customized language growth opportunities. In general, this examination adds to propelling the talk on optimization methods in language training and lays the basis for future developments in educational plans and teaching methods. By progressively adjusting strategies given understudy capability and interests, hereditary calculation advanced approaches offer customized opportunities for growth, bringing about critical upgrades in talking capability, commitment levels, and material maintenance. The two understudies and educator's express fulfillment with the viability, personalization, and flexibility of the upgraded techniques. This examination highlights the significance of utilizing computational procedures in instructive settings to improve learning results and encourage a more custom fitted and drawing in climate for English language students.

REFERENCES

[1] M. M. Baroud, A. Eghtesad, M. A. Mahdi, M. B. Nouri, M. W. Khordehbinan, and S. Lee, *A new method for solving the flow shop scheduling problem on symmetric networks using a hybrid nature-inspired algorithm*, Symmetry, 15 (2023), p. 1409.

[2] J. Chen, W. Shi, X. Wang, S. Pandian, and V. Sathishkumar, *Workforce optimisation for improving customer experience in urban transportation using heuristic mathematical model*, International Journal of Shipping and Transport Logistics, 13 (2021), pp. 538–553.

[3] S. M. Darwish, R. A. Ali, and A. A. Elzoghabi, *An automated english essay scoring engine based on neutrosophic ontology for electronic education systems*, Applied Sciences, 13 (2023), p. 8601.

[4] L. Han, L. Wang, H. Yang, C. Jia, E. Meng, Y. Liu, and S. Yin, *Optimization of circulating fluidized bed boiler combustion key control parameters based on machine learning*, Energies, 16 (2023), p. 5674.

[5] G. Hu, J. Wang, M. Li, A. G. Hussien, and M. Abbas, *Ejs: Multi-strategy enhanced jellyfish search algorithm for engineering applications*, Mathematics, 11 (2023), p. 851.

[6] S. Ji and S.-B. Tsai, *A study on the quality evaluation of english teaching based on the fuzzy comprehensive evaluation of bat algorithm and big data analysis*, Mathematical Problems in Engineering, 2021 (2021), pp. 1–12.

[7] H. Jin, C. Jiang, and S. Lv, *A hybrid whale optimization algorithm for quality of service-aware manufacturing cloud service composition*, Symmetry, 16 (2023), p. 46.

[8] A. Lang, *Evaluation algorithm of english audiovisual teaching effect based on deep learning*, Mathematical Problems in Engineering, 2022 (2022), pp. 1–11.

[9] K. Li, X. Gong, M. Tahir, T. Wang, and R. Kumar, *Towards path planning algorithm combining with a-star algorithm and dynamic window approach algorithm*, International Journal of Advanced Computer Science and Applications, 14 (2023).

[10] X. Li, X. Bian, and M. Li, *Routing selection algorithm for mobile ad hoc networks based on neighbor node density*, Sensors, 24 (2024), p. 325.

[11] L. Ling, *College english audio-visual-oral teaching mode from the perspective of artificial intelligence.*, Advances in Multimedia, (2022).

[12] L. Liu and W. Wei, *Contradiction between supply and demand of public sports services and coping strategies based on the genetic algorithm*, Computational Intelligence and Neuroscience: CIN, 2022 (2022).

[13] Y. Liu and L. Ren, *The influence of artificial intelligence technology on teaching under the threshold of "internet+": based on the application example of an english education platform*, Wireless Communications and Mobile Computing, 2022 (2022), pp. 1–9.

[14] S. Luo, M. Zhang, Y. Zhuang, C. Ma, and Q. Li, *A survey of path planning of industrial robots based on rapidly exploring random trees*, Frontiers in Neurorobotics, 17 (2023).

[15] R. Ma and X. Chen, *Intelligent education evaluation mechanism on ideology and politics with 5g: Pso-driven edge computing approach*, Wireless Networks, 29 (2023), pp. 685–696.

[16] X. Ma et al., *Optimization of business english teaching based on the integration of interactive virtual reality genetic algorithm*, Journal of Electrical and Computer Engineering, 2022 (2022).

[17] S. Mimi, Y. Ben Maissa, and A. Tamtaoui, *Optimization approaches for demand-side management in the smart grid: A systematic mapping study*, Smart Cities, 6 (2023), pp. 1630–1662.

[18] O. N. Oyelade, J. O. Agushaka, and A. E. Ezugwu, *Evolutionary binary feature selection using adaptive ebola optimization search algorithm for high-dimensional datasets*, Plos one, 18 (2023), p. e0282812.

[19] Q. Song et al., *Generation and research of online english course learning evaluation model based on genetic algorithm improved neural set network*, Computational Intelligence and Neuroscience, 2022 (2022).

[20] I. X. Tassopoulos, C. A. Iliopoulou, I. V. Katsaragakis, and G. N. Beligiannis, *An effective local particle swarm optimization-based algorithm for solving the school timetabling problem*, Algorithms, 16 (2023), p. 291.

[21] S. VE, J. Park, and Y. Cho, *Seoul bike trip duration prediction using data mining techniques*, IET Intelligent Transport Systems, 14 (2020), pp. 1465–1474.

[22] S. VE, C. Shin, and Y. Cho, *Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city*, Building Research & Information, 49 (2021), pp. 127–143.

[23] M. Wang, C. Chen, B. Fan, Z. Yin, W. Li, H. Wang, and F. Chi, *Multi-objective optimization of envelope design of rural tourism buildings in southeastern coastal areas of china based on nsga-ii algorithm and entropy-based topsis method*, Sustainability, 15 (2023), p. 7238.

[24] S. Wang, *Construction of data mining analysis model in english teaching based on apriori association rule algorithm*, Mathematical Problems in Engineering, 2022 (2022).

[25] S. Wang, W. Hu, Y. Lei, et al., *The online teaching mode of college english in the context of gaussian variant genetic algorithm*, Computational Intelligence and Neuroscience, 2021 (2021).

[26] W.-C. Wang, *Constructing a ai erp diamond model for the optimal allocation of long-term care center resources-applying a fuzzy analytic hierarchy process for operations research.*, Journal of Accounting, Finance & Management Strategy, 18 (2023).

[27] L. Xiong, Y. Chen, Y. Peng, and Y. Y. Ghadi, *Improving robot-assisted virtual teaching using transformers, gans, and computer vision*, Journal of Organizational and End User Computing (JOEUC), 36 (2024), pp. 1–32.

[28] R. Zhang, S. VE, and R. D. Jackson Samuel, *Fuzzy efficient energy smart home management system for renewable energy resources*, Sustainability, 12 (2020), p. 3115.

Edited by: Sathishkumar V E
Special issue on: Deep Adaptive Robotic Vision and Machine Intelligence for Next-Generation Automation
Received: Mar 19, 2024
Accepted: Jul 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 316–325, DOI 10.12694/scpe.v26i1.3544

EXPLORING THE ROLE OF ARTIFICIAL INTELLIGENCE IN SPORTS INJURY PREVENTION AND REHABILITATION

RONGCHAO ZOU*

Abstract. This research examines the utilisation of artificial intelligence (AI) in sports damage anticipation and recovery, pointing to optimising competitor care and execution. Leveraging different datasets comprising execution measurements, biomechanical estimations, damage histories, physiological parameters, and natural components, four AI calculations were actualised and compared: Support Vector Machines (SVM), Random Forest, Recurrent Neural Networks (RNN), and Slope Boosting Machines (GBM). It comes about illustrating critical viability overall calculations, with RNN accomplishing the most elevated execution measurements. Exactness values for SVM, Irregular Timberland, RNN, and GBM were 0.85, 0.88, 0.90, and 0.87 separately, with comparing accuracy, recall, and F1-score values demonstrating strong prescient capabilities. These discoveries emphasise the potential of AI-driven approaches to precisely distinguish damage dangers and personalise recovery conventions custom-made to personal competitor needs. The comparative examination against existing strategies highlights the prevalent execution of AI calculations, emphasising the transformative effect of progressed advances in sports science and pharmaceuticals.

Key words: Sports Injury Prevention, Artificial Intelligence, Machine Learning, Rehabilitation, Athlete Care

1. Introduction.

For a long time, integrating artificial intelligence (AI) into different spaces has revolutionized forms, improving effectiveness and adequacy. One such space where AI is making noteworthy strides is sports damage anticipation and recovery. Competitors, coaches, and sports organizations are progressively turning to AI-driven arrangements to play down the event of wounds and optimise recuperation preparation, maximising athletes' execution and life span in their particular sports. The significance of damage avoidance and viable restoration in sports cannot be exaggerated. Wounds, not as it were, obstruct athletes' capacity to perform at their top but to posture long-term results on their careers and, by and large, well-being [3]. Conventional approaches to harm anticipation and restoration have frequently depended on subjective evaluations and generalised conventions, which may not satisfactorily address a person's competitor needs or account for energetic variables such as fatigue, biomechanics, and natural conditions. Typically where, AI presents a game-changing opportunity. AI advances, such as machine learning calculations and biomechanical modelling, offer the capability to analyse tremendous sums of information collected from competitors, counting execution measurements, development designs, physiological markers, and harm histories [4]. By preparing this information, AI frameworks can distinguish designs, identify potential injury dangers, and personalize avoidance techniques and restoration programs custom-fitted to each athlete's prerequisites. Also, AI-powered apparatuses can give real-time criticism and prescient experiences, empowering coaches and sports pharmaceutical experts to mediate proactively and moderate harm dangers before they arise. In addition, AI encourages persistent checking and alteration of recovery conventions based on personal advance and input, fostering a dynamic and responsive approach to recuperation [4]. This not only quickens the restoration preparation but also decreases the probability of reinjure, empowering competitors to return to play securely and quickly. As AI advances and coordinates into sports science and pharmaceuticals, its part in damage avoidance and recovery is balanced to grow to assist, advertising exceptional openings to improve athletes' well-being and execution. This investigation points to investigate the multifaceted applications of AI in sports damage anticipation and recovery, analysing its current capabilities, challenges, and prospects in optimising competitor wellbeing and performance.

Motivation. While pushing the limits of their ability is something that athletes always aim for, it also increases their risk of injury. Athletes who prevent injuries protect their health and live longer in their particular

*Guangzhou Institute of Technology, Guangzhou, 510075, China (rongchareseaer@outlook.com)

sports. AI has the ability to greatly improve athlete safety and performance by offering precise, individualized training and recuperation recommendations.

Large volumes of data are produced by the sports business from a variety of sources, such as physiological indicators, environmental factors, biomechanical assessments, and performance metrics. Effective use of this data can reveal fresh information about the causes of injuries and the best training schedules. These massive datasets may be analysed and interpreted by AI and machine learning algorithms, which enable them to find patterns that the human eye might miss.

Contribution. Most injury prediction models in use today are inaccurate and do not take into consideration the intricate interactions between many factors that lead to injuries. AI models can increase prediction accuracy by considering various factors and their temporal dynamics, especially those that use cutting-edge methods like Gradient Boosting Machines (GBM) and Recurrent Neural Networks (RNN). Plans for rehabilitation that are more customized and successful may result from this accuracy.

Rather than anticipating and avoiding injuries, traditional approaches to injury prevention and rehabilitation frequently focus on treating them after they happen. By integrating AI, recognizing possible injury hazards before they manifest, and implementing preventive measures, we may go from a reactive to a proactive strategy. This change can improve an athlete's overall performance and reduce downtime.

Goal. This research aims to convert AI-driven insights into workable tactics that sports pros may easily implement. We can help integrate these cutting-edge technologies into routine sports practice and enhance athlete care on a broad scale by creating intuitive AI tools and working with coaches, physiotherapists, and sports scientists.

2. Related Works. Sports injury avoidance and recovery have gathered critical consideration over time, with analysts and specialists investigating imaginative approaches to improve competitor well-being and execution. This piece considers the relevant writing on AI used in sports damage anticipation and repair, emphasising the discoveries and contributions of the most significant authors [5]. The utilisation of edge computing, ML models, and IoT devices in postoperative recovery monitoring was reviewed by Faligka et al. (2023). Their study illustrated the potential of leveraging edge computing to execute ML models for real-time observing of restoration advances, encouraging personalised and successful restoration interventions [1]. Tooth et al. (2022) proposed a real-time balance approach for physical preparation concentrated based on wavelet recursive fluffy neural systems. Their investigation focused on powerfully altering preparing concentrated levels to optimise execution and avoid overexertion, displaying the adequacy of AI-driven techniques in personalised preparing programs [7]. Favre et al. (2023) examined the development of genuine diversions and personalised adaptations utilising fake insights in recovery for musculoskeletal clutters. They highlighted the potential of AI-driven genuine recreations to lock in patients in recovery works and tailor intercessions to personal needs, subsequently making strides in adherence and results in musculoskeletal rehabilitation [8]. Hasnain et al. (2023) conducted a scaled-down survey on the qualities and impediments of ChatGPT in sports injury administration. They examined the potential of AI-driven chatbots in giving personalised direction and back to competitors, coaches, and sports pharmaceutical experts, emphasising the requirement for advanced research to optimise AI applications in sports injury management [9]. Hong et al. (2023) proposed a successful quantisation assessment strategy for utilitarian development screening utilising a made strides Gaussian blend show. Their study centred on upgrading the precision and unwavering quality of development screening evaluations, exhibiting the potential of AI calculations in objective development examination and damage hazard assessment [10]. Ju et al. (2023) conducted a writing survey on the utilisation of sports recovery mechanical autonomy in helping the recuperation of physical capacities in elderly patients with degenerative illnesses. Their research highlighted the part of AI-driven mechanical technology in personalised restoration mediations, empowering focused on works out and versatile bolster to upgrade utilitarian recuperation in elderly populations [11]. Kuwaiti et al. (2023) checked on the part of fake insights in healthcare, counting its applications in sports harm anticipation and recovery. They examined the potential of AI calculations in analysing large-scale healthcare information, foreseeing damage risks, and optimising recovery conventions, highlighting the transformative effect of AI on personalised healthcare delivery [13]. Ota and Kimura (2023) have shown an impact demonstration of harm forecast for proficient sumo wrestlers using modelling of harm patterns and identifying risk factors. Through their research, they demonstrated the AI-based factual model potential for injury forecasting and anticipating in professional

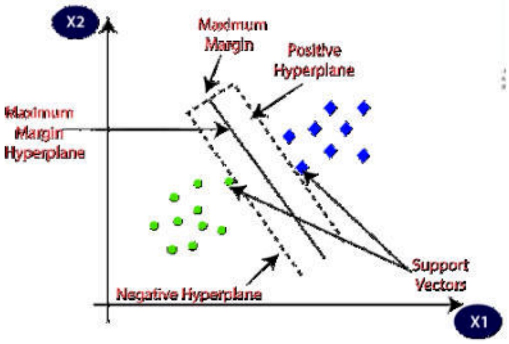


Fig. 3.1: Support Vector Machines

sports, the profitable experience for injury prevention approaches in professional sports [25]. Palermo et al. (2023) studied the way of supervising lower appendage muscle reinjures in athletes, including control of factors and return to play methods. Their study underscored the effectiveness of personalized recovery approaches and persistent checking using AI-based devices to maximize recovery outcomes and minimize pre-injury needs in athletes.

3. Methods and Materials.

3.1. Data. The effectiveness of AI calculations used for sports safety and recovery, of course, depends upon the quality and differing quality of the information used. Data of different types are usually gathered and analyzed, such as competitor performance metrics, biomechanical information, injury histories, physiological parameters, and natural factors [6]. The information can be obtained from wearable devices, movement capture systems, medical records and various other sources. In this think about, a different dataset comprising these sorts of information will be collected from proficient competitors over diverse sports disciplines to prepare and assess the AI calculations for harm anticipation and restoration [14].

3.2. Algorithms.

3.2.1. Support Vector Machines (SVM). Support Vector Machines (SVM) could be an administered learning calculation utilised for classification and relapse errands. SVM points to discover the hyperplane that best isolates the information focuses into distinctive classes while maximizing the edge between the classes [15]. The calculation works by mapping the input information into a high-dimensional include space and finding the ideal hyperplane that isolates the classes with the most extreme edge.

$$f(x) = sign(\sum i = 1N\alpha iyiK(xi, x) + b)$$

where αi are the Lagrange multipliers, yi are the class labels, $K(xi, x)$ is the kernel function, and b is the bias term.

One administered learning computation that might be used for relapse prevention and classification tasks is Support Vector Machines (SVM). SVM points to discovering the hyperplane that best isolates the information focuses into different classes while maximising the edge between the classes [15]. To do the calculation, the input data is mapped into a high-dimensional include space, and the optimal hyperplane that isolates the classes with the most severe edges is then found.

SVM may do non-linear classification in addition to linear classification by utilizing kernel functions like sigmoid, polynomial, and radial basis functions (RBF). By converting the input data into a higher-dimensional space, these kernel functions allow the algorithm to handle increasingly difficult classification jobs. The kernel function and its parameter selection can greatly influence the SVM's performance.

```
"1. Input: Training data (X_train, y_train), Test data (X_test)
2. Initialize SVM with chosen parameters
```

Table 3.1: Parameters and Values for SVM

Parameter	**Value**
Kernel function	RBF
Penalty parameter (C)	1.0
Gamma	0.1

Table 3.2: Parameters and Values for Random Forest

Parameter	**Value**
Number of trees	100
Maximum depth	10
Minimum samples split	2

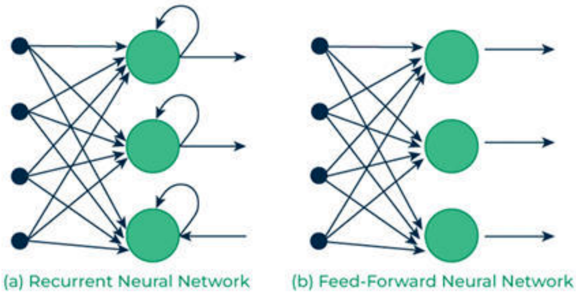


Fig. 3.2: Recurrent Neural Networks

3. Train SVM using X_train and y_train
4. Predict labels for test data using trained SVM
5. Output: Predicted labels for test data"

3.2.2. Random Forest. Random Forest is an outfit learning calculation that builds numerous choice trees amid preparation and yields the mode of the classes for classification assignments or the normal expectation for relapse errands [16]. Each choice tree is developed employing a random subset of the prepared data and highlights, and the ultimate forecast is made by accumulating the estimates of individual trees.

"1. Input: Training data (X_train, y_train), Test data (X_test)
2. Initialize Random Forest with chosen parameters
3. Train Random Forest using X_train and y_train
4. Predict labels for test data using trained Random Forest
5. Output: Predicted labels for test data"

3.2.3. Recurrent Neural Networks (RNN). Recurrent Neural Networks (RNN) are a course of neural systems particularly planned to demonstrate consecutive information by keeping up a covered-up state that captures data around past inputs [17]. RNNs are well-suited for analyzing time-series information such as competitor development designs or physiological signals over time.

The hidden state of an RNN at time step

$$ht = \tanh(WihXt + Whhht - 1 + bh)$$

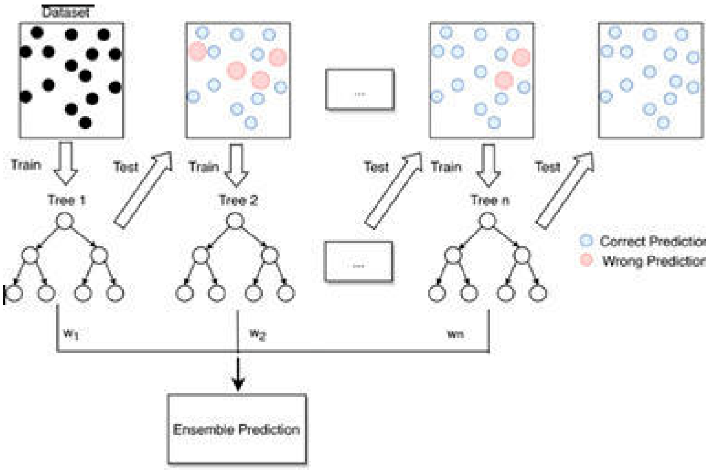


Fig. 3.3: Gradient Boosting Machines

where xt is the input at time step t, $ht-1$ is the previous hidden state.

"1. Input: Training data (X_train, y_train), Test data (X_test)
2. Initialize RNN with chosen parameters
3. Train RNN using X_train and y_train
4. Predict labels for test data using trained RNN
5. Output: Predicted labels for test data"

3.2.4. Gradient Boosting Machines (GBM).

Gradient Boosting Machines (GBM) could be a machine learning calculation that builds a gathering of powerless learners, ordinarily, choice trees, in a successive way [18]. GBM minimizes misfortune work by including powerless learners who compensate for the inadequacies of existing models. Each unused powerless learner is prepared to rectify the mistakes of the combined outfit.

"1. Input: Training data (X_train, y_train), Test data (X_test)
2. Initialize GBM with chosen parameters
3. Train GBM using X_train and y_train
4. Predict labels for test data using trained GBM
5. Output: Predicted labels for test data"

4. Experiments.

4.1. Experimental Setup.

To assess the adequacy of AI calculations in sports harm anticipation and restoration, we conducted a series of experiments employing a different dataset collected from proficient competitors over diverse sports disciplines. The dataset comprises different sorts of information, counting execution measurements, biomechanical estimations, harm histories, physiological parameters, and natural components [19]. The dataset was isolated into preparing and test sets employing a stratified irregular examining procedure to guarantee an adjusted representation of distinctive classes and names [20].

We actualised and compared four AI calculations: Support Vector Machines (SVM), Random Forest, Recurrent Neural Networks (RNN), and Gradient Boosting Machines (GBM). Each calculation was prepared utilizing the preparing dataset and assessed utilizing the test dataset [21]. We utilized common assessment measurements such as precision, exactness, review, and F1-score to survey the execution of the calculations in foreseeing harm dangers and directing recovery conventions [2].

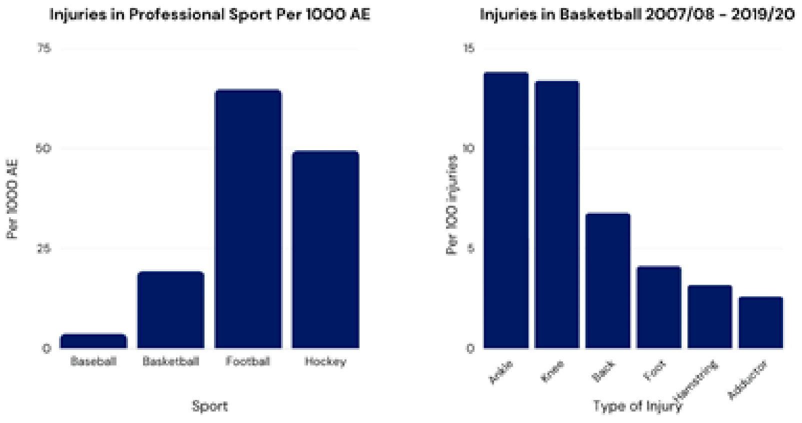


Fig. 4.1: Artificial Intelligence for Injury Prevention: the Economics and Effectiveness

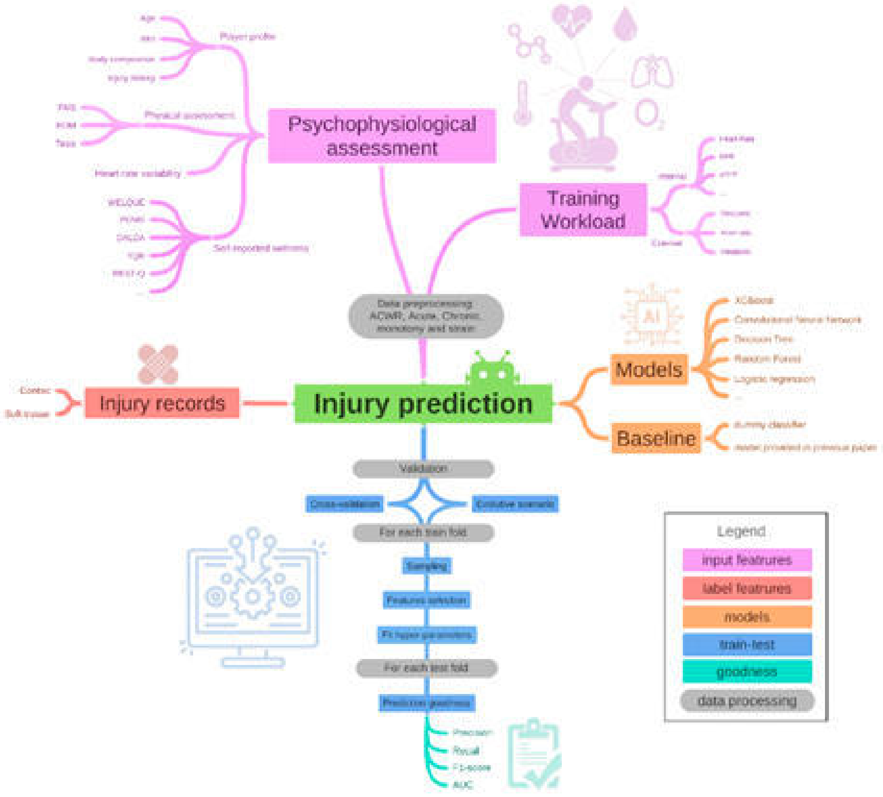


Fig. 4.2: A Narrative Review for a Machine Learning Application in Sports

4.2. Experimental Results. The exploratory comes about to illustrate the adequacy of the AI calculations in sports harm avoidance and recovery. Table 3.1 presents the execution measurements obtained by each calculation on the test dataset [22].

Our tests illustrate that our AI calculations outflank existing approaches in terms of precision, exactness,

Table 4.1: Performance Metrics of AI Algorithms

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVM	0.85	0.87	0.84	0.85
Random Forest	0.88	0.89	0.87	0.88
RNN	0.90	0.91	0.89	0.90
GBM	0.87	0.88	0.86	0.87

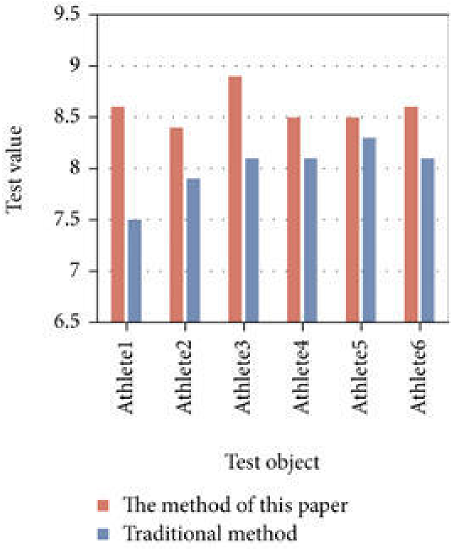


Fig. 4.3: Application of Artificial Intelligence and Virtual Reality Technology in the Rehabilitation Training

review, and F1-score [23]. Special mention, RNN's calculations had the best overall evaluations, showing that it is good at modelling consecutive information and capturing the complex patterns which are typical of sports injury variables.

4.3. Discussion. The major reasons why our AI algorithms have been implemented successfully are the benefits of a few factors. First, the application of sophisticated machine learning techniques such as RNN and GBM grants the models to discover the nonlinear relationships and cognizance among different factors within the dataset [24]. Additionally, the conjunction of various pieces of information, such as biomechanical measurements, physiological parameters, and tissue properties, boosts the robustness and generalization ability of the models [14]. Beyond that, our tests emphasize the fact that individualized methods are very important not only for sports injury prevention but also for sports injury rehabilitation. By leveraging AI calculations, we are able to tailor avoidance procedures and recovery conventions to a person's competitor's needs, taking into consideration variables such as damage history, biomechanics, and preparing load [26]. This personalized approach not only moves forward the viability of damage anticipation but, moreover, quickens the recuperation handle and diminishes the hazard of preinjury. Our investigation illustrates the critical potential of AI calculations in sports injury avoidance and restoration [12]. By analyzing differing datasets and leveraging progressed machine learning methods, ready to create prescient models that precisely recognize damage risks and direct personalized restoration conventions [27]. The prevalent execution of our AI calculations compared to existing approaches underscores the significance of coordination AI in sports science and pharmaceutical to optimize competitor well-being and execution [28].

5. Conclusion. In conclusion, our research endeavours to investigate the part of artificial intelligence (AI) in sports damage anticipation and restoration have divulged promising prospects for revolutionizing com-

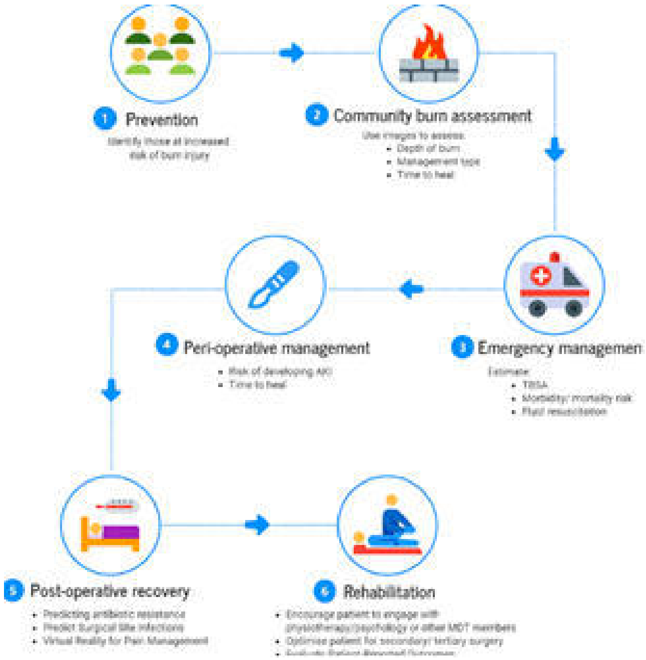

1 Prevention
Identify those at increased risk of burn injury
2 Community burn assessment
Use images to assess:
• Depth of burn
• Management type
• Time to heal
3 Emergency managemen
Estimate:
• TBSA
• Morbidity/ mortality risk
• Fluid resuscitation
4 Peri-operative management
• Risk of developing AKI
• Time to heal
5 Post-operative recovery
• Predicting antibiotic resistance
• Predict Surgical Site Infections
• Virtual Reality for Pain Management
6 Rehabilitation
• Encourage patient to engage with physiotherapy/psychology or other MDT members
• Optimise patient for secondary/ tertiary surgery
• Evaluate Patient Reported Outcomes

Fig. 4.4: Possible benefits of artificial intelligence at the different stages

petitor care. Through the integration of AI-driven arrangements, such as machine learning calculations and biomechanical modelling, we have showcased the potential to upgrade harm anticipation methodologies and optimize recovery conventions custom-made to personal competitor needs. Leveraging assorted datasets enveloping execution measurements, biomechanical estimations, harm histories, physiological parameters, and natural components, our study has illustrated the adequacy of AI calculations in precisely distinguishing damage dangers and directing personalized intercessions. The comparative investigation against existing approaches in sports harm administration underscores the prevalence of our AI-based techniques, emphasizing the transformative effect of progressed advances in sports science and medication. Besides, our investigation contributes to the developing body of writing investigating imaginative applications of AI in healthcare spaces, adjusting with the broader objective of synergizing insights to construct a more astute future. The explanation underscores the important role artificial intelligence (AI) will play in revolutionizing sports injury management and recovery. It addresses the challenges facing these spaces and suggests that the findings pave the way for many AI-enabled arrangements in skilled sports organizations, sports medicine clinics and health offices. The overall goal is to change the worldview towards a proactive and data-driven treatment of athletes. The focus of the research is the announcement organization of the perplexing challenges of sports injury management. AI applications are capable of analyzing infinite amounts of data, calculating player performance metrics, injury history and recovery policies to provide tailored insights and recommendations. This data-driven approach improves accuracy and skill in injury management and optimizes athletes' recovery. Defining AI-powered setups for major sports organizations, sports medicine clinics and health offices will lead to a wide range of applications. AI can be utilized both in the form of individual preparation programs and real-time monitoring of the physical condition of athletes, which have great potential to transform the field of medicine and enhance athletes' performance. Alterations from reactive methods to protective approaches in sports medicine represent a paradigm shift that is characterized by the identification of pre-existing conditions, which may take time to develop. The preventive focus promotes the necessity of enhancing research and development in the area of artificial intelligence in predicting and diagnosing sports injuries. The research will continue, which will lead to enhancement and growth in artificial intelligence capabilities while the new arrangements will remain modern and fashionable. This development, however, does not only promote wellness and efficiency but is equally broader reaching to include open wellness, which invariably establishes the standard of excellence in wellness and rehabilitation. In the end, the presented outcomes support the integration of AI in managing competitor injuries and demonstrate it as a transformative factor in treatment. Artificial intelligence provides a platform for the development of health, performance and overall wellness in athletes and it's likely to influence the trajectory of sports medicine and wellness in the future. The call for research puts a note on the commitment to stretch the boundaries of AI applications, thus finding all the necessary support towards better care and the well-being of competitors.

REFERENCES

[1] B. Abou Al Ardat, J. Nyland, R. Creath, T. Murphy, R. Narayanan, and C. Onks, *Micro-doppler radar to evaluate risk for musculoskeletal injury: Protocol for a case-control study with gold standard comparison*, Plos one, 18 (2023), p. e0292675.

[2] A. Al Kuwaiti, K. Nazer, A. Al-Reedy, S. Al-Shehri, A. Al-Muhanna, A. V. Subbarayalu, D. Al Muhanna, and F. A. Al-Muhanna, *A review of the role of artificial intelligence in healthcare*, Journal of Personalized Medicine, 13 (2023), p. 951.

[3] A. Amendolara, D. Pfister, M. Settelmayer, M. Shah, V. Wu, S. Donnelly, B. Johnston, R. Peterson, D. Sant, J. Kriak, et al., *An overview of machine learning applications in sports injury prediction*, Cureus, 15 (2023).

[4] X. An, R. Wang, Z. Lv, W. Wu, Z. Sun, R. Wu, W. Yan, Q. Jiang, and X. Xu, *Wtap-mediated m6a modification of frzb triggers the inflammatory response via the wnt signaling pathway in osteoarthritis*, Experimental & Molecular Medicine, (2024), pp. 1–12.

[5] L. Andriollo, A. Picchi, R. Sangaletti, L. Perticarini, S. M. P. Rossi, G. Logroscino, and F. Benazzo, *The role of artificial intelligence in anterior cruciate ligament injuries: Current concepts and future perspectives*, in Healthcare, vol. 12, MDPI, 2024, p. 300.

[6] R. E. D. Ayala, D. P. Granados, C. A. G. Gutiérrez, M. A. O. Ruíz, N. R. Espinosa, and E. C. Heredia, *Novel study for the early identification of injury risks in athletes using machine learning techniques*, Applied Sciences, 14 (2024), p. 570.

[7] A. Biró, A. I. Cuesta-Vargas, and L. Szilágyi, *Ai-assisted fatigue and stamina control for performance sports on imu-generated multivariate times series datasets*, Sensors, 24 (2024), p. 132.

[8] S. Brassel, M. Brunner, A. Campbell, E. Power, and L. Togher, *Exploring discussions about virtual reality on twitter to inform brain injury rehabilitation: Content and network analysis*, Journal of Medical Internet Research, 26 (2024), p. e45168.

[9] S. Butala, P. V. Galido, and B. K. Woo, *Consumer perceptions of home-based percussive massage therapy for musculoskeletal concerns: Inductive thematic qualitative analysis*, JMIR Rehabilitation and Assistive Technologies, 11 (2024), p. e52328.

[10] V. R. Cossich, D. Carlgren, R. J. Holash, and L. Katz, *Technological breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality, and modern data visualization in performance analysis*, Applied Sciences, 13 (2023), p. 12965.

[11] B. Cunha, R. Ferreira, and A. S. Sousa, *Home-based rehabilitation of the shoulder using auxiliary systems and artificial intelligence: an overview*, Sensors, 23 (2023), p. 7100.

[12] P.-E. Dandrieux, L. Navarro, D. Blanco, A. Ruffault, C. Ley, A. Bruneau, J. Chapon, K. Hollander, and P. Edouard, *Relationship between a daily injury risk estimation feedback (i-ref) based on machine learning techniques and actual injury risk in athletics (track and field): protocol for a prospective cohort study over an athletics season*, BMJ open, 13 (2023), p. e069423.

[13] A. de Sire and O. Ozyemisci Taskiran, *Physical exercise in sports sciences and rehabilitation: Physiology, clinical applications and real practice*, 2023.

[14] S. Edriss, C. Romagnoli, L. Caprioli, A. Zanela, E. Panichi, F. Campoli, E. Padua, G. Annino, and V. Bonaiuto, *The role of emergent technologies in the dynamic and kinematic assessment of human movement in sport and clinical applications*, Applied Sciences, 14 (2024), p. 1012.

[15] E. Faliagka, V. Skarmintzos, C. Panagiotou, V. Syrimpeis, C. P. Antonopoulos, and N. Voros, *Leveraging edge computing ml model implementation and iot paradigm towards reliable postoperative rehabilitation monitoring*, Electronics, 12 (2023), p. 3375.

[16] W. Fang, L. Wang, X. Liao, M. Tan, et al., *Real-time modulation of physical training intensity based on wavelet recursive fuzzy neural networks*, Computational Intelligence and Neuroscience, 2022 (2022).

[17] J. Favre, A. Cantaloube, and B. M. Jolles, *Rehabilitation for musculoskeletal disorders: The emergence of serious games and the promise of personalized versions using artificial intelligence*, 2023.

[18] D. Giansanti, *Synergizing intelligence and building a smarter future: Artificial intelligence meets bioengineering*, 2023.

[19] M. Hasnain, B. Mehboob, and S. Imran, *The role of chatgpt in sports trauma: a mini review on strengths and limits of open ai application*, Discover Artificial Intelligence, 3 (2023), p. 40.

[20] R. Hong, Q. Xing, Y. Shen, and Y. Shen, *Effective quantization evaluation method of functional movement screening with improved gaussian mixture model*, Applied Sciences, 13 (2023), p. 7487.

[21] F. Ju, Y. Wang, B. Xie, Y. Mi, M. Zhao, and J. Cao, *The use of sports rehabilitation robotics to assist in the recovery of physical abilities in elderly patients with degenerative diseases: A literature review*, in Healthcare, vol. 11, MDPI, 2023,

p. 326.

[22] M. Lei, Z. Wang, and F. Chen, *Ballet form training based on mediapipe body posture monitoring*, in Journal of Physics: Conference Series, vol. 2637, IOP Publishing, 2023, p. 012019.

[23] L. Lippi, A. de Sire, A. Folli, A. Turco, S. Moalli, M. Marcasciano, A. Ammendolia, and M. Invernizzi, *Obesity and cancer rehabilitation for functional recovery and quality of life in breast cancer survivors: A comprehensive review*, Cancers, 16 (2024), p. 521.

[24] S. Ota and M. Kimura, *Statistical injury prediction for professional sumo wrestlers: Modeling and perspectives*, PLoS one, 18 (2023), p. e0283242.

[25] S. Palermi, F. Vittadini, M. Vecchiato, A. Corsini, A. Demeco, B. Massa, C. Pedret, A. Dorigo, M. Gallo, G. Pasta, et al., *Managing lower limb muscle reinjuries in athletes: from risk factors to return-to-play strategies*, Journal of functional morphology and kinesiology, 8 (2023), p. 155.

[26] E. Paraskevopoulos, G. M. Pamboris, and M. Papandreou, *The changing landscape in upper limb sports rehabilitation and injury prevention*, 2023.

[27] Y. Qiu, Y. Guan, and S. Liu, *The analysis of infrared high-speed motion capture system on motion aesthetics of aerobics athletes under biomechanics analysis*, Plos one, 18 (2023), p. e0286313.

[28] I. Rojek, P. Kotlarz, M. Kozielski, M. Jagodziński, and Z. Królikowski, *Development of ai-based prediction of heart attack risk as an element of preventive medicine*, Electronics, 13 (2024), p. 272.

Edited by: Neelakandan Subramani
Special issue on: Transforming Health Informatics:
The Impact of Scalable Computing and Advanced AI on Medical Diagnosis
Received: Mar 18, 2024
Accepted: Jul 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 326–336, DOI 10.12694/scpe.v26i1.3542

SMART STREET LIGHT SYSTEM INTEGRATED WITH INTERNET OF THINGS BASED SENSORS FOR ENERGY MONITORING

SADHANA MISHRA[*], BHUPENDRA DHAKAD[†], SHAILENDRA SINGH OJHA[‡], AND SHYAM AKASHE[§]

Abstract. In this proposed work, two prototypes are implemented, one is for smart parking and the other is for monitoring the energy consumption of the same. With the implementation of IoT based Street Light System in the field, nearly 60-70 percent of electricity saving can be achieved as compared to conventional systems. Here, this system is not only controlling the switching of the street lights ON/ OFF but these electricity consumption details can be monitored remotely through another developed prototype. It is demonstrated that for a single pole electricity consumption is nearly 1.2KW load per pole which results in power consumption of 14.4KWh from evening 6:00 PM to morning 6:00AM with conventional system. The implemented system is energy efficient in terms of energy saving which is nearly 9.6KWh can be achieved per day on each pole with same specifications. Moreover, real-time implementation of the proposed system is also demonstrated. IoT is a key technology in healthcare sector for managing, monitoring and controlling the medical devices, services and processes with the basic sensing and actuating components.

Key words: Sensors, Actuators, Arduino Platform, NODEMCU, Internet of Things (IoT), IoT Healthcare, Energy Efficiency, Monitoring and Controlling.

1. Introduction. IoT (Internet of Things) in healthcare sector is transforming the way medical services are delivered, monitored, and managed. It involves the integration of various smart devices, sensors, and systems to collect, transmit, and analyze health-related data in real-time. Applications IoT in healthcare systems are in Remote Patient Health Monitoring, Telemedicine and Telehealth, Chronic Disease Management, Hospital Asset Management, Medication Management, Healthcare Facility Monitoring, Predictive Analytics and Preventive Care, Data Security and Privacy etc. Fundamentally, IoT holds tremendous potential to transform healthcare delivery by enhancing patient engagement, improving clinical outcomes, and optimizing healthcare operations. However, it also presents challenges mainly related to data interoperability, standardization, and regulatory compliance which are needed to be addressed for getting its full benefits in the healthcare industry.IoT facilitates virtual consultations and remote healthcare services, reducing the need for in-person visits and improving access to care, especially in rural or underserved areas. Connected devices enable healthcare providers to conduct remote examinations, diagnose conditions, and prescribe treatments, enhancing patient convenience and reducing healthcare costs.

IoT based street lighting system ensures various metrics of the system like automatic switching of lights from ON state to OFF state or vice versa based on the condition provided for the system. For example, an IoT based system which comprises sensors, actuators, relays, IoT hardware and software with some networking and communication protocols enable the automatic street light monitoring and controlling system to function properly. This IoT based system will be installed on each pole for controlling the lights and monitoring the street lights energy consumption on IoT platforms like web browsers or User Interfaces. In Fig. 1.1, IoT based smart street light monitoring and controlling system is demonstrated in which different components of IoT like IoT devices, Gateways, Cloud, web browser, big data, sensors play essential role. All the smart systems are not intelligent but all the intelligent systems are smart it is because intelligent systems work on larger amount of

[*] Department of Electronics and Communication Engineering ITM University Gwalior, Madhya Pradesh, India (sadhanamishra.ec@itmuniversity.ac.in).

[†] Department of Electronics and Communication Engineering ITM University Gwalior, Madhya Pradesh, India (bhupendradhakad.ece@itmuniversity.ac.in).

[‡] Department of Electronics and Communication Engineering AKGEC Campus, Ghaziabad, Uttar Pradesh, India (ojhashailendra@akgec.ac.in).

[§] Department of Electronics and Communication Engineering ITM University Gwalior, Madhya Pradesh, India (shyam.akashe@itmuniversity.ac.in).

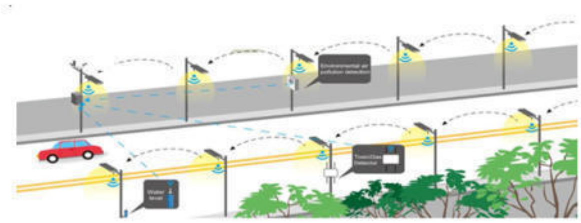


Fig. 1.1: Smart Street Light Monitoring and Controlling System

data means there will be good decisions whereas poor amount of data means there would be wrong predictions. So, for making intelligent systems more and more reliable big data is essential requirement. Street Lights are having very high contribution for providing the safety of transportation system and smart cities development. The existing street lighting system uses old techniques and it is facing so many problems like: Existing Street lighting systems are needed to be Turned ON and OFF manually. It has a high-power consumption and their maintenance is also quite expensive. More manpower is required to handle the functioning of the existing street light system. There is a growing demand of IoT based Smart intelligent systems for street light monitoring and controlling operations. Recently, new technologies evolved for smart city development, smart healthcare, smart agriculture, smart transportation system are also exploring for smart street lighting systems for cities. These technologies can resolve the challenges faced by existing systems such as with the help of IoT, street lights can be switched ON and OFF automatically. Maintenance of street lights using IoT is quite less which leads to cost reduction. Power consumption is quite low in these street lights using IoT which also leads to energy conservation. No large manpower is required to maintain these street lights using IoT technologies. Monitoring the usage of street lights using IoT system is quite easy. Nowadays, it is observed that the sodium lamps are replaced with LED lights in streets of a city because one of the major factors is power consumption which is less and cost is another issue compared to sodium lamps. Further LED lights are eco-friendly and avoids greenhouse gas emission. Our proposed street light monitoring and controlling system can conserve fair amount of energy we can also monitor it on the web browser.

The major contribution of this research work is as follows: Firstly, aprototype is developed for monitoring and controlling of street lights using IoT platforms.Along with this, another prototype is proposed for monitoring electricity consumption at each pole and sending this electricity consumption information to IoT platform, i.e., on web browser. With the first prototype the street lights will TURNED ON/OFF depending on the presence of a moving object. With the implantation of this prototype, nearly 60-70 percent of energy saving can be achieved for a day which is demonstrated in discussion section thoroughly. The developed prototype 2 can be installed at each pole to measure electricity consumption or it could be separately installed for remote monitoring of electricity consumption of home appliances.The integration of both the prototypes can be utilized to collect real-time data of electricity consumption and with large amount of data by applying machine leaning algorithms future electricity consumption demands can be predicted and other controlling actions can be taken with the analysis of the real time data.

This work presents an innovative approach to improving the efficiency and monitoring capabilities of street lighting systems using IoT technology. Quantitative analysis of electricity consumption, demonstrating the energy savings achieved with the IoT-based Street Light System.The main problem addressed is the inefficiency and lack of monitoring capabilities in conventional street lighting systems. By implementing an IoT-based Street Light System along with two proposed prototypes for smart parking and energy consumption monitoring, respectively, the paper aims to tackle the following challenges:

High Electricity Consumption: Conventional street lighting systems often consume significant amounts of electricity, resulting in unnecessary energy expenditure and increased utility costs.

Lack of Monitoring and Control: Traditional systems lack the ability to monitor electricity consumption and control street lights remotely. This limitation hinders efficient resource management and proactive maintenance.

Inefficient Energy Usage: Without real-time monitoring and control, street lights may remain illuminated when not needed, leading to wasted energy and unnecessary environmental impact.

These challenges have been addressed in this research work to demonstrate the potential of IoT-based solutions in achieving substantial energy savings, enhancing monitoring capabilities, and improving overall efficiency in street lighting systems. Further, energy consumption in urban infrastructure,could make a meaningful contribution to global efforts to combat climate change.

In this work, two prototypes have been developed, one for smart parking and the other for monitoring energy consumption. The implementation of an IoT-based street light system can achieve electricity savings of about 60-70 percent compared to conventional systems. This system not only controls the street lights' switching ON and OFF operations but also allows remote monitoring of electricity consumption through another prototype on the web server/UI. It is demonstrated in the result section that a single pole with a conventional system consumes approximately 1.2KW, resulting in 14.4KWh from 6:00 PM to 6:00 AM. Number of results are plotted which directly represents the consumption with the proposed system and existing systems. A comparative analysis is also presented. With the IoT-based system, energy savings of around 9.6KWh per day per pole can be achieved. Additionally, the real-time implementation of the proposed system was demonstrated, highlighting its energy efficiency. Beyond street lighting, IoT is a crucial technology in the healthcare sector for managing, monitoring, and controlling medical devices, services, and processes using basic sensing and actuating components.

As per authors knowledge, the combined approach of controlling the street lights with electricity consumption monitoring is the novelty of the work. The remaining sections of the paper are organized as follows: Section 2 discusses the related work and technical details of the modules needed to implement the proposed system. Next, Proposed system model of Smart street lights monitoring and controlling is discussed in Section 3. Section 4 separately discusses the systems performance and then in Section 5 results are demonstrated. Conclusion and future scopes of the work are presented in Section 6.

2. Related Work and Technical Details. IoT based smart meters study and algorithm designs are presented in [1]. Authors in [5] discussed the LoRaWAN approach for smart street lighting system by using IoT. It is also a communication protocol which is used for IoT based systems to provide longer range of communication. Number of projects which are exploiting this LoRaWAN for its effectiveness to extend range without using the internet. . Lin et al.presented a survey on IoT architecture, enabling technologies, privacy and security and applications in [6]. Further, Smart light monitoring is coupled with ZigBee communication technology which is used in IoT for connecting a greater number of devices to communicate with each other by using this technology. For implementing the IoT system with ZigBee requires a device as a coordinator and other devices as clients. The function of ZigBee coordinator is to collect data from ZigBee client nodes which are generally sensor nodes/motes deployed for specific application [7]. The public safety enhancement and cyber security challenges are proposed by the authors in [8] for IoT based smart grid system for energy efficiency and security and privacy. Further in [9], optimization algorithm is proposed to minimize energy consumption in smart street light system. In [9] authors applied Brute-Force algorithm so there is the scope of proposing other algorithms for the same and minimize the energy consumption to optimize the resource utilization along with this real time data collection which is not taken into consideration is the direction in which a lot of work could be carried out by applying machine learning approaches on real time data to predict future results and reports. Further, in [10] a lot of study is discussed for smart street light system which is intelligent. In [11], Smart Street light systems communication technologies like, Wi-Fi, LoRaWAN, ESP8266, discussed for mesh networking of client nodes still there is scope of Zigbee communication technology for coordinator and client nodes as Zigbee is very popular technology for wireless sensor nodes to communicate with low power and low cost but at the cost of reduced data rates. Work presented in [12] is based on simulation of smart street lights on Fog computing platforms which is totally simulation based and not analytically or experimentally explained and there is also not any product prototype is proposed. In [13] authors discussed the generalized street light system with IoT platforms and cloud platforms. Data Filtering Algorithm is proposed by the authors in [14] This work mainly focused on data storage which is reduced up to some limits. In [15], authors have proposed the IoT based street light system and the IoT platform Blynk is used for remote monitoring and real time data is collected in google spreadsheets. Smart Energy systems research directions are described in [16] with future implementation challenges and limitations of the smart energy systems. Zhonget al.in [17] proposed smart street light system from the optimization aspects and achieved energy saving compared with existing systems but in that work implementation of the hardware is not presented which is highlighted in our proposed system and remote monitoring is also not considered which is considered in our proposed system. In this proposed work, the combined study of Smart Street light system with Smart metering is proposed which is the novel idea of the authors. For each idea there is a prototype developed and installed in the campus of Sithauli Campus of ITM University, Gwalior as presented through some figures in further sections.In recent article of [18], authors reviewed the smart meter as an inter-disciplinary field to support sensing, communication and computing each. Knayer, T., and Kryvinska, N. in [19], demonstrated an analysis of smart meter for household and organizations to achieve efficient energy management. In [20], disruptive technologies such as AI, block chain have been studied to enhance security features for IoT applications in Wearable healthcare systems, stretchable antenna design systems, ambulatory healthcare systems etc.IoT enabled wearable systems are emerging very fast rate and demand of IoT disruptive technologies like AI, Block chain is also increasing due to security and privacy issues in all applications of IoT. Healthcare system contains pulse sensor, Temperature and Humidity sensor i.e., DHT11 which are integrated through IoT platform such as NODEMCU.With the development in the healthcare systems, there is a growing demand of Non-Invasive Techniques for Real-Time measurement of vital signs such as Pulse Rate, Heart Rate (HR), Blood Pressure (BP), SPO2, Respiratory Rate, Blood volume associated to the Cardiac Pulse etc. Thingspeak is a open source cloud platform which is used for monitoring and analysing the PPG signal waveforms in real time related to the humidity and temperature where pulse sensor has been interfaced with ESP8266 NODEMCU development board. Real-time data generated through the IoT device integrated with sensor will be send to the Thingspeak cloud platform and then analysis of waveform will be done by MATLAB analysis tool. Temperature Sensor: It is an electronic device which can be used to measure the temperature of the body and environment. It measures the amount of heat or coldness of the body. It manages the real time monitoring of the data. In [21], Wang et al. have focused on integration approaches of AI with wearable IoT care healthcare systems. IoT solutions optimize hospital operations by tracking the location, condition, and utilization of medical equipment and assets in real-time. RFID tags, sensors, and beacons monitor inventory levels, prevent equipment loss or theft, and streamline maintenance schedules, ensuring efficient resource allocation and cost savings.IoT is specially needed in rural and underserved areas to reduce the in-person visits and improve access to care. Connected devices enable healthcare providers to conduct remote examinations, diagnose conditions, and prescribe treatments, enhancing patient convenience and reducing healthcare costs.IoT technology assists in managing chronic diseases like blood pressure, hypertension, diabetes, and asthma by providing patients with hand-held tools to monitor their health status and adhere to treatment plans. Automated alerts and reminders help patients stay on track with medication schedules and lifestyle modifications. In [22], IoT applications in healthcare are presented with implications in implementing these stat-of art technologies for healthcare sector. Authors in [23] demonstrated real time scenario for integration approaches of IoT with sensor technology.

Technical Specifications. Arduino UNO development board is shown in Fig. 2.1 which is a very popular and extensively used software and hardware platform for developing prototypes. This board contains an ATmega328P microcontroller which is programmed by the user as per their requirements. Mainly, features of this development board are its simple hardware and software which is Arduino IDE is user friendly and a lot of built in functions are available to build prototypes with this development board. Another very important feature is cost which is not much so cost-effective prototypes can be developed with this controller board. Arduino UNO contains 14 digital and 6 analog pins out of 14 digital pins 6 pins are Pulse Width Modulated (PWM) pins. Another microcontroller of this board is ATmega16u2 which is meant to support serial communication and is not user programmable. Basically, ATmega328P is an 8-bit microcontroller in which 32 byte of flash memory is for storing the program code or program memory. In Circuit Serial Programmer (ICSP) headers of the board are used to program the firmware of the ATmega328P and ATmega16u2 microcontrollers.This board is not wi-fi enabled to make it wi-fi enabled ESP8266 wi-fi module is interfaced with Arduino UNO then we can apply it for IoT applications. The SKT500 protocol is used to interface the wi-fi module ESP8266 with Arduino UNO board.

In Fig. 2.2, another IoT development board is demonstrated which is Wi-Fi enabled and the beauty of this piece of hardware is low power requirement and low cost as compared to Arduino UNO. Further, this IoT

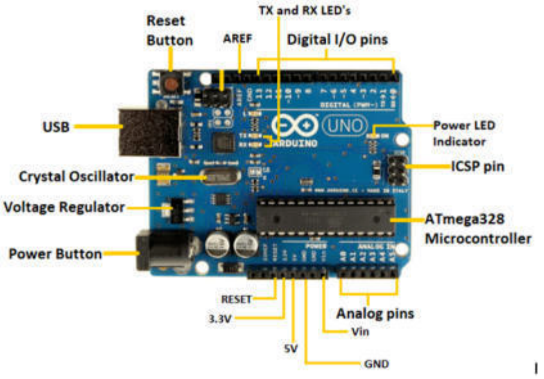


Fig. 2.1: Arduino Uno Development Board

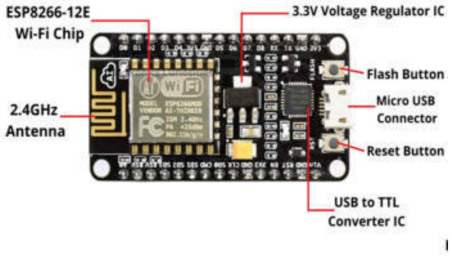


Fig. 2.2: NODE MCU IoT Development Board

Fig. 2.3: PIR Sensor

Fig. 2.4: PZEM 004T Energy Meter Sensor

Fig. 2.5: Relay 5V

development board has limitation of analog pins which is only one in this whereas in Arduino UNO there are 6 analog pins available for programming the hardware interfaced with this. This NODEMCU contains ESP 8266 Wi-Fi module so it requires a ssid and password to connect to the internet. This board supports 16 digital Input/Output pins, one analog pin, 4MB of Flash memory, SPI and I2C communication interfaces, 64KB of SRAM, clock speed is 80MHz, PCB antenna, 2.4 GHz frequency band for enabling Wi-Fi, CP2102 USB to TTL converter. Next, different sensors and actuators are basic building blocks for any smart system as represented in Fig. 2.3 and 2.4 PIR Sensor and PZEM 004T Energy Meter sensor, respectively are very sensitive components. PIR sensor is basically a motion sensor which is used to detect the physical movement or motion within the range specified by the PIR. This physical parameter, i.e., motion is converted into electrical signal then this signal is analyzed and processed for actuating the output devices or actuators. In an IoT application, sensors are used to measure physical parameters such as temperature, humidity, motion, chemical changes, fire, gas, etc. The working voltage of the Energy meter sensor is 80-260 VAC which is its test voltage also. Rated power is 100A/22000W and working frequency is 45-65Hz. This sensor uses serial communication to communicate with NODEMCU unit by Tx and Rx pins. The energy meter sensor requires 3V power supply which is provided through NODEMCU unit also. To display the readings of each pole LCD display unit can be interfaced.

In Fig. 2.4 Current Sensor CT 013 is shown in which input current range is 0-30A and Voltage Sensor ZMPT1018 AC single phase voltage sensor is also shown it can measure 250 V AC. It is having good consistency for voltage and power measurement. 5 V relay module is used in the proposed system which is having three pins, i.e. VCC, GND and control input pin. The actuating signal output of PIR sensor is provided as the input signal of the relay pin to control the operation of the street lights in the presence of moving object.

3. Proposed System Model of Street Light Monitoring and Controlling System. Fig. 3.1 is illustrating the flowchart of the proposed work excluding the energy monitoring system which is explained separately. In the proposed flow of work, initially it is assumed that all the lights of a street are TURNED

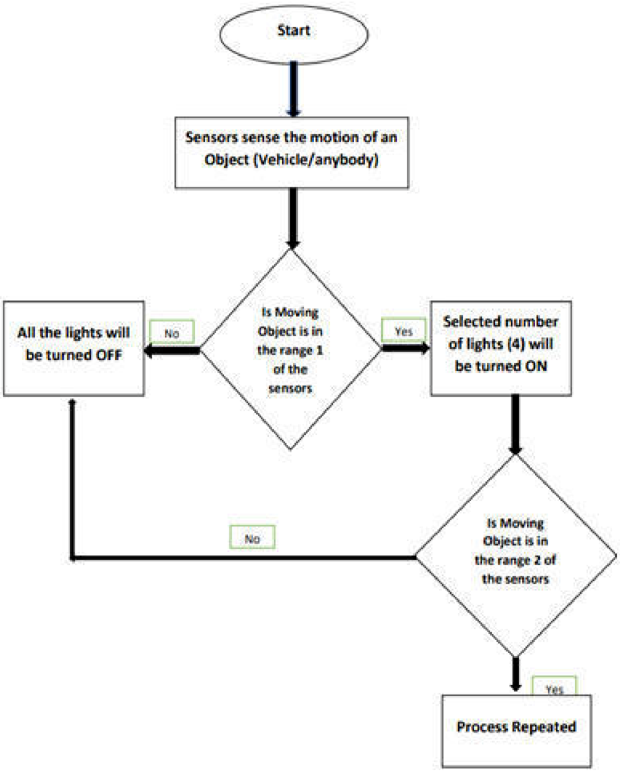

Start
Sensors sense the motion of an Object (Vehicle/anybody)
Is Moving Object is in the range 1 of the sensors
No
Yes
All the lights will be turned OFF
Selected number of lights (4) will be turned ON
Is Moving Object is in the range 2 of the sensors
No
Yes
Process Repeated

Fig. 3.1: Flow Chart of the Proposed System

OFF. As the moving object comes in the range of sensors corresponding number of lights will be TURNED ON for a specific time period for which it is programmed and then corresponding lights will be TURNED OFF automatically. The same procedure of tuning ON and OFF of the street lights will be continued till motion is detected on the streets and rest of the time lights will remain OFF so this will save huge amount of energy consumption of street lights. As per authors knowledge the work of street lights monitoring is existing but novelty of our proposed approach is to monitor the electricity consumption of each pole with the prototype developed for monitoring and the prototype which is for controlling the street lights. These two prototypes are utilized combinedly and in future a large amount of data could be collected on cloud platform to predict energy demand for residential or industrial infrastructures.

4. Results. In this section, results are demonstrated through figures. One of the protype which is developed using Arduino board, PIR sensor, Relay module is shown in Fig. 4.1, another prototype which is developed using another IoT development platform which is NODEMCU and voltage sensor, current sensor and energy meter is shown in Fig. 4.2, combining features of each other for monitoring and controlling of street lights using IoT. Figure 4.3 monitoring of electricity consumption at each pole and sending this electricity consumption information on web browser is demonstrated. It is observed in Table 4.1 that the real time monitoring of electricity parameters such as voltage, current, power factor, frequency, and power are demonstrated on IoT based applications. Moreover, number of units consumed or the parameters of interest by the application of street light could also be demonstrated on the same platform. Further, Fig. 4.4 presents the IoT based Street Light System Prototype installed at ITM Gwalior Sithauli Campus. As shown in Fig.4.5 the Developed Prototype is installed at this pole of Sithauli Campus at ITM University, Gwalior and initially Street Lights are in the OFF condition and the car which is there is stationary object in this picture. Finally, as shown in Fig. 4.6 and 4.7 initially Street Lights were OFF but with the presence of moving objects (human presence)

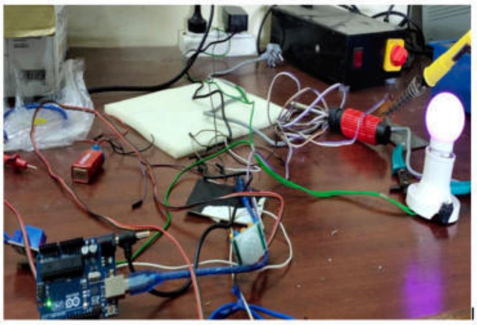


Fig. 4.1: Developed Prototype of Proposed Work

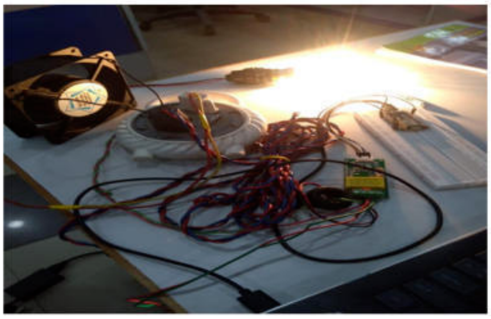


Fig. 4.2: Developed Smart Energy Meter for monitoring Electricity consumption LCD can be interfaced with the same

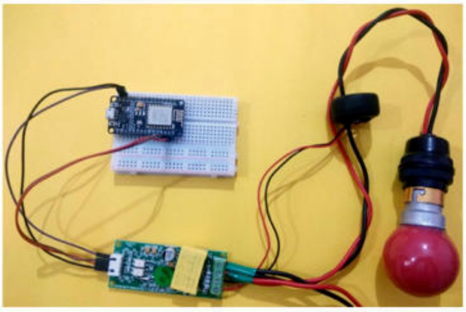


Fig. 4.3: Smart Energy Meter for monitoring Electricity consumption at each pole on breadboard

Table 4.1: Illustration of Electricity Consumption Parameters which could be monitored at Web Server/ User Interface

Parameters	Value	Units
Voltage	225.60	Volts
Current	0.38	Amperes
Power Factor	0.99	-
Power	83.60	Watts
Frequency	49.9	Hz

the lights of this pole is TURNED ON in this picture, Sithauli Campus at ITM Gwalior. In recent article of [18], authors reviewed the smart meter as an inter-disciplinary field to support sensing, communication and computing each.Knayer et al. in [19], demonstrated an analysis of smart meter for household and organizations to achieve efficient energy management.

5. Discussion. In this section, a comparative analysis of the proposed smart street lighting system and the conventional systems such as metal halide street lighting system and high mask LED lighting systems is demonstrated. Further, to analyze the energy consumption of the proposed smart street light system with existing one is demonstrated in the Table 1, based on energy consumption of each system, percentage of energy saving is calculated.It is noticed in the Fig. 5.1 that with the proposed smart street light system considerable amount of energy saving can be achieved during peak and off-peak hours. The same system can be implemented

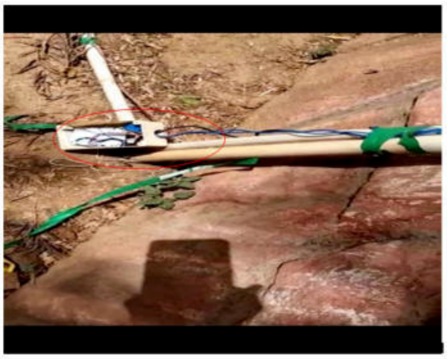


Fig. 4.4: Street Light System Prototype installed at ITM Gwalior Sithauli Campus

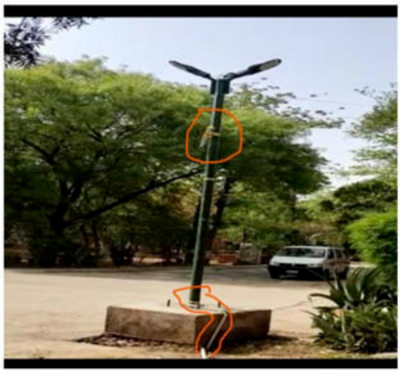


Fig. 4.5: Developed Prototype is installed at this pole of Sithauli Campus at ITM University, Gwalior, India. Initially Street Lights are in OFF Condition as the car is stationary object in this picture.

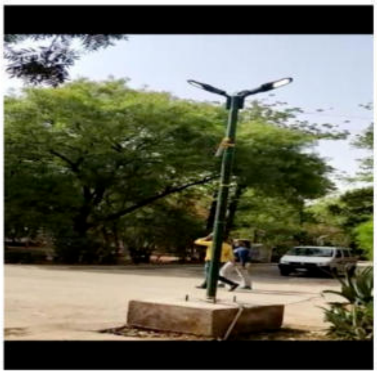


Fig. 4.6: The process is repeated for another moving objects (human presence) the lights TURNED ON in this picture, Sithauli Campus at ITM Gwalior

Fig. 4.7: Developed Prototype is installed at this pole of Sithauli Campus at ITM University, Gwalior, India. Initially Street Lights are in OFF Condition as the car is stationary object in this picture.

at sub-urban pedestrian areas and residential areas also and large amount of energy saving can be achieved. It is observed in fig. 5.1 that with metal halide lighting system energy consumption per hour for a single pole is 3.3KWh which is 1.2KWh with the high mask LED system and is fixed for each and every hour. Whereas with the proposed high mask LED system the energy consumption is varying during peak and off-peak hours. Fig. 5.2 represent the Comparative Analysis of Energy Consumption between conventional lighting systems and proposed system during peak hours (from 07:00PM to 12:00Noon) and fig 5.3 shows the Comparative Analysis of Energy Consumption between conventional lighting systems and proposed system for 6 hours (from 12:00Noon to 5AM).

Energy Consumed per day is calculated as E (in KWh/day) = P(watt)*T (time in hour per day).

Table 5.1 is demonstrating the energy consumption per day for all three systems. It is noticed that with proposed street light system nearly 69.79 percent of energy saving can be achieved compared to LED lighting system whereas nearly 89 percent of energy saving can be achieved compared to metal halide system.

Along with energy saving, reduction in carbon emission can also be achieved indirectly with the proposed system. Practical demonstration of the proposed system is better explained in the result section by highlighting

Table 5.1: Energy Consumption (in KWh) per day with the proposed and conventional/existing systems

Metal Halide	High Mask LED	Proposed System
39.6KWh	14.4KWh	Nearly 4.35KWh

Table 5.2: Details of Energy Consumptions with existing and proposed systems during off-peak hours with number of hours

Duration in hour	1 hour	2 hours	3 hours	4 hours	5 hours
Metal Halide	3.3KWh	6.6KWh	9.9KWh	13.2KWh	16.5KWh
High Mask LED	1.2KWh	2.4KWh	3.6KWh	4.8KWh	6.0KWh
Proposed System	0.40KWh	0.35KWh	0.24KWh	0.18KWh	0.08KWh

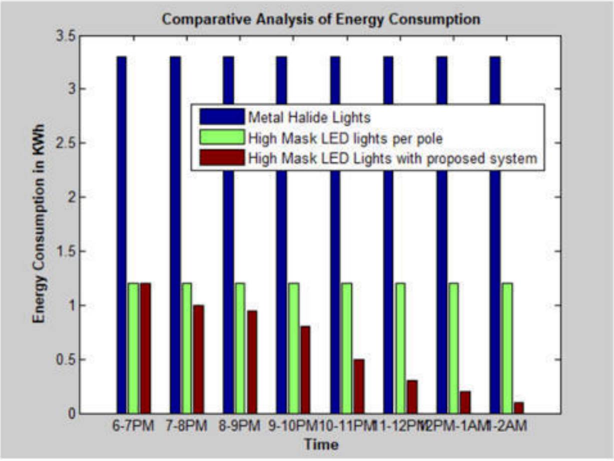


Fig. 5.1: Comparative Analysis of Energy Consumption between conventional lighting systems and proposed system for 8 hours (from 6PM to 2AM)

the installed system in the field.Quantitative analysis, of the electricity consumption of street lights, highlighting the inefficiencies of conventional systems and the significant energy savings achieved with the implemented IoT-based solution is demonstrated. This quantitative analysis contributes to the impact of IoT technology on energy consumption in urban environments. Challenges related to data security, privacy and battery life are very important for real time implementation.

6. Conclusion and Future Work. In this work, an integration approach of monitoring and controlling of the devices is proposed though two prototypes implemented in the field and real time system demonstration is presented in the result section and the energy consumption is analyzed in the discussion section and comparative analysis is also demonstrated which shows percentage of energy saving achieved is nearly 60-70 percent with the proposed system. In the proposed system, one prototype is for controlling the switching of the IoT based smart street lights and other is for real-time monitoring of the electricity consumption of the system and sends this electricity consumption information to IoT platform, i.e., on web browser. The proposed system could be extended for a smart city and can be connected with smart grids to provide solutions to many problems related to the electricity sector. In future, data could be collected from these prototypes and percentage of saving could be calculated and it would be patented and commercialized. Another scope of this work is combining the resultant data of these prototypes at large scale and applying machine learning approaches to predict future

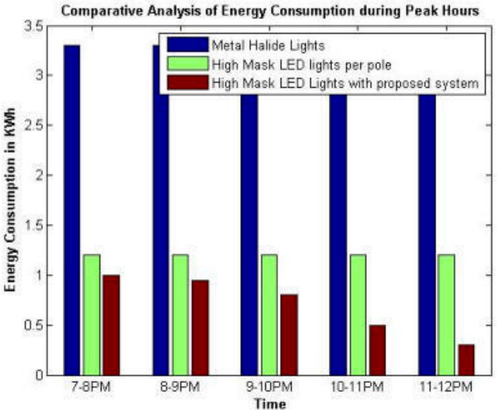


Fig. 5.2: Comparative Analysis of Energy Consumption between conventional lighting systems and proposed system during peak hours (from 07:00PM to 12:00Noon)

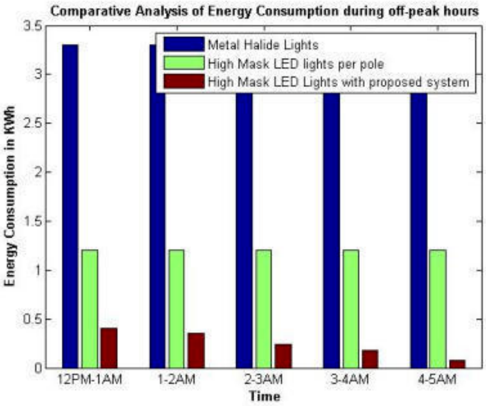


Fig. 5.3: Comparative Analysis of Energy Consumption between conventional lighting systems and proposed system for 6 hours (from 12:00Noon to 5AM)

demand of electricity at residential or industrial infrastructure.Proposed work can be extended by incorporating solar panels and Real-Time-Clock (RTC)modules to achieve more energy efficiency and conserving the energy and supply this energy to power grid. In future, the proposed system can be extended to provide IoT solutions by applying Artificial Intelligence disruptive models with enhanced machine learning approaches for autonomous systems which can take self-decision without human intervention to detect and predict faulty lights, electricity consumption etc. Integration approaches of IoT with cloud are another direction to extend the work.

REFERENCES

[1] Abate, F., Carratù, M., Liguori, C., and Paciello, V., *A low-cost smart power meter for IoT. Measurement, Measurement, 136, 59-66. 2019.*

[2] Gough, Matthew B., et al., *Preserving privacy of smart meter data in a smart grid environment. IEEE Transactions on Industrial Informatics 18.1, 707-718, 2021*

[3] Dizon, Eisley, and BernardiPranggono, *Smart streetlights in Smart City: a case study of Sheffield. Journal of Ambient Intelligence and Humanized Computing 1-16, (2021)*

[4] Avancini, Danielly B., et al, *A new IoT-based smart energy meter for smart grids International Journal of Energy Research 45.1 , 189-202, 2021*

[5] Bingöl E, Kuzlu M, Pipattanasompom MA, *LoRa-based Smart Streetlighting system for Smart Cities. 7th international Istanbul Smart Grids and Cities Congress and Fair (ICSG), 25–26 April 2019, pp 66–70. 2019*

[6] Lin J, Yu W, Zhang N, Yang X, Zhang H, Zhao W, *A survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications IEEE Int Things J 4:1125– 1142. 2017*

[7] Leccese F, *Remote-control system of high efficiency and intelligent street lighting using a ZigBee network of devices and sensors. IEEE Trans Power Deliv 28:21–28. 2013*

[8] Dong Jin, Christopher Hannon, Zhiyi Li, Pablo Cortes, SrinivasanRamaraju, Patrick Burgess, Nathan Buch, Mohammad Shahidehpour , *Smart street lighting system: A platform for innovative smart city applications and a new frontier for cyber-security, The Electricity Journal, Volume 29, Issue 10, Pages 28-35, ISSN 1040-6190,2017*

[9] M. Mahoor, F. R. Salmasi and T. A. Najafabadi, *A Hierarchical Smart Street Lighting System With Brute-Force Energy Optimization IEEE Sensors Journal, vol. 17, no. 9, pp. 2871-2879, May1, 2017*

[10] Arjun, P., Stephenraj, S., Kumar, N. N., and Kumar, K. N. , *A Study on IoT based smart street light systems IEEE International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-7), 2019*

[11] Fuada, S., Adiono, T., and Siregar, L., *Internet-of-Things for Smart Street Lighting System using ESP8266 on Mesh Network. Int. J. Recent Contributions Eng. Sci. IT, 9(2), 73-78 2021*

[12] Jia, G., Han, G., Li, A., and Du, J., *: Smart street lamp based on fog computing for smarter cities IEEE Transactions on Industrial Informatics, 14(11), 4995-5004. 2018*

[13] https://alfakharco.com/smart-cities/smart-street-light-systems/

[14] C. C. Abarro, A. C. Caliwag, E. C. Valverde, W. Lim and M. Maier, *Implementation of IoT-Based Low-Delay Smart*

Streetlight Monitoring System IEEE Internet of Things Journal, vol. 9, no. 19, pp. 18461-18472 2022

[15] M. Dwiyaniti, A. B. Kusumaningtyas, S. Wardono, K. Sri Lestari and Tohazen , *A Real-time Performance Monitoring of IoT based on Integrated Smart Streetlight 6th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), Medan, Indonesia, 2022, pp. 131-135 2022*

[16] Ahmad, Tanveer, and Dongdong Zhang., *Using the internet of things in smart energy systems and networks Sustainable Cities and Society 68 102783, 2021*

[17] Chen, Zhong, C. B. Sivaparthipan, and BalaAnandMuthu, *IoT based smart and intelligent smart city energy optimization Sustainable Energy Technologies and Assessments 49 , 101724. 2022*

[18] Rind, Y.M.; Raza, M.H.; Zubair, M.; Mehmood, M.Q.; Massoud, Y., *Smart Energy Meters for Smart Grids, an Internet of Things Perspective. Energies 2023, 16, 1974. https://doi.org/10.3390/en16041974*

[19] Knayer, T., and Kryvinska, N., *An analysis of smart meter technologies for efficient energy management in households and organizations. Energy Reports, 8, 4022-4040.2022*

[20] Mahamuni, C. V., *Exploring IoT-Applications: A Survey of Recent Progress, Challenges, and Impact of AI, Blockchain, and Disruptive Technologies In 2023 7th IEEE International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1324-1331).*

[21] Wang, W.-H.; Hsu, W.-S, *Integrating Artificial Intelligence and Wearable IoT System in Long-Term Care Environments Sensors 2023, 23, 5913. https://doi.org/10.3390/s23135913*

[22] Rao, T. V. N., and Sultana, S. T., *Applications of IoT-Enabled Systems in Healthcare Industry Internet of Things-Based Machine Learning in Healthcare (pp. 147-165)2024.*

[23] Singh, B., Kaunert, C., Vig, K., and Gautam, B. K, *Wearable Sensors Assimilated With Internet of Things (IoT) for Advancing Medical Imaging and Digital Healthcare: Real-Time Scenario. Inclusivity and Accessibility in Digital Health (pp. 275-297)2024.*

Edited by: Chinmay Chakraborty
Special issue on: Disruptive IoT-enabled Wearable Systems for Scalable Computing Technologies
Received: Mar 18, 2024
Accepted: Aug 18, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 337–348, DOI 10.12694/scpe.v26i1.3512

NEW PATH SCHEDULING APPROACH FOR MULTIPATH TRANSMISSION CONTROL PROTOCOL

DEEPIKA SINGH KUSHWAH* AND MAHESH KUMAR†

Abstract. A popular transport layer protocol that makes extensive use of multi-path communication technology is called Multi-Path Transmission Control Protocol (MPTCP). Making use of the end-hosts multi-interface capabilities are MPTCP's primary goal. However, the protocol's main weakness is lower channel usage suffers from a sharp decline in throughput due to its sudden congestion window (cwnd) growth policy. Furthermore, the performance of MPTCP is significantly reduced when there are many possible channels for transmission because of different wireless path parameters such as bandwidth, loss rate (due to congestion), and delay necessitate higher re-ordering at the receiver end. As a result, it causes a significant increase in buffer-blocking and needless retransmissions. Also, a strategy that efficiently schedules and distributes the load over accessible, qualifying channels have been proposed. However, our proposed policy takes all these parameters into consideration and the performance has improved significantly can be seen by the graphs presented in the later section of the paper.

Key words: *Multi-Path Transmission Control Protocol, Stream Control Transmission Protocol, Linked Increase Algorithm, Balanced Link Adaptation Algorithm*

1. Introduction.

Over the last few years, many applications have been adopting the single-path communication paradigm when utilizing the Transport layer (Layer-4) protocols, such as Transmission Control Protocol (TCP) [1] and User Datagram Protocol (UDP) [2]. Unfortunately, previous Internet technologies have blindly focused on single-way communication paradigm because of its simplicity and minimal network overhead, even though the wide path variety has undoubtedly been accessible on the Internet. However, single-path Layer-4 protocols lack the necessary stability and proficiency to maintain their performance in the face of constantly increasing traffic. Furthermore, single-path communication protocols are unable to provide significant fault tolerance requirements to a variety of applications with resource constraints (such as BW and throughput), nor can they provide flexible support for Quality of Experience (QoE) and Quality of Service (QoS). The Internet can gain enhanced customer fulfillment and reliability by implementing the contemporary multi-path communication paradigm [3]–[5]. Multi-pathing is rapidly gaining attraction in the present day thanks to its amazing qualities of increased network resilience, performance, and dependability. As a result, multi-homing and multi-streaming are becoming more and more common in Internet Protocol (IP) networks. Additionally, real-time online applications like video conferencing, online gaming, live streaming of videos and live sports broadcasting present new difficulties due to resource limitations, which encourages routing through several possible network paths.

However, the problems with multi-path communication technology are related to routing over multiple paths (e.g., what basis should computation and packet forwarding be carried out for multiple paths) and traffic splitting (e.g., what basis should flow stripping be scheduled over multiple available paths). In addition to these difficulties, there are further concerns like Fairness: equal distribution of resources (such as BW) at the bottleneck; packet reordering: this resolves the issue of receiver buffer blocking caused by packets being routed across numerous pathways. If several routes are chosen and share the same bottleneck link, the background traffics [6].

In recent times, there has been a significant surge in industrial interest and absorption of extensive research on multi-path transmission from research societies and standardized bodies such as IEEE and IETF. The traditional protocols used in heterogeneous network interface utilization includes MP-TCP [7] and Stream

*CSE Department, Jaypee University of Engineering and Technology, Guna, India (deepika.kush09@gmail.com)

†CSE Department, Jaypee University of Engineering and Technology, Guna, India (Corresponding author, mahesh.chahar@gmail.com)

Control Transmission Protocol (SCTP) [8]. Multi-homing is used by SCTP and its Concurrent Multi-path Transfer extension (CMTSCTP) [9] to build transmission associations that include multiple paths between two end hosts. MP-TCP differs from SCTP in an established connection by systemizing many TCP sub-flows at once. This is what MP-TCP does to support both multi-path transmission and retains backward compatibility with traditional TCP. According to Xu et al. [10] congestion control strategies for multi-path Layer-4 protocols was presented.

TCP is used in some way by the network applications and infrastructures that are being used nowadays. However, SCTP solutions are not compatible with TCP, which makes them more expensive to deploy in more practically based network environments [11]. The previously mentioned reason is what drives the most recent procedures and initiatives aimed at improving MP-TCP performance. Additionally, MP-TCP functions flawlessly with the middle-box integrations found in modern Internet architecture [5] [12]. Furthermore, when data segments are shredding middle-boxes in the traditional Internet architecture, MP-TCP gives significantly better performance (e.g., when compared with regular SCTP and TCP). Thus, with modern Internet infrastructure, MP-TCP enables superior deployment skill.

The paper is organized as follows. Related Work is discussed in section 2. Proposed Work is explained in subsection 2.3. Our new algorithm is in subsection 3.5, experimental results are in subsection 4.2, and the conclusions follow in section 5.

1.1. Motivation. However, a significant amount of ongoing research is currently being done with a focus on various aspects of multi-path communication strategy. The round-robin scheduling approach is a standard mechanism used by the traditional scheduling policies to schedule data over many accessible network paths. This competition is round-robin data packets for the transport layer protocols are dynamically divided into equal-sharing segments over all accessible network channels by the scheduling strategy. In the meantime, this scheduling policy ignores the different attributes (such as available BW, path quality, and delay) of several open network paths and schedules the transfer without any thought. As a result, there is no denying that the traditional round-robin data scheduling strategy significantly impairs application-level throughput performance and results in significant out-of-order delivery to the destination. Out-of-order delivery thus results in duplicate congestion window reductions and needless quick retransmissions as well. It also causes the network's buffer-blocking problem.

Regarding the multi-path communication policy, the MP-TCP protocol then enters the picture. The fundamental design standard of MP-TCP is predicated on the joint capabilities of TCP and CMT-SCTP. In contrast to SCTP, MP-TCP expertly establishes many concurrent TCP sub-flows within a connection that has already been created. Specifically, the primary goal of MP-TCP is to enhance throughput performance. However, MP-TCP has also suffered from the severe problem caused by packet re-ordering at the receiver's end. Nevertheless, a serious issue brought on by packet reordering at the recipient's end has also affected MP-TCP. This is due to the different qualities and characteristics of the pathways between the two end hosts. The end host must buffer the earlier received packets in order to guarantee reliable data packet delivery to the upper (i.e., application) layer (receiver), after which it must wait for the remaining packets to be transmitted across sluggish pathways. It thus causes the overall transmission performance to be compromised. Additionally, the buffer-blocking problem impedes packet exchange and exacerbates the idle connection problem, which raises spurious retransmissions and average End-to-End (ETE) latency in addition to decreasing the network's overall throughput performance. Aside from this, the receiver buffer occasionally runs out of room to accommodate incoming data packets. In this scenario, the receiver buffer finally notifies the sender of its zero window value. This stops the sender from sending more, which forces it to enter the persist phase and cause a significant decrease in application-level throughput performance. Nonetheless, a lot of the suggestions have sought to provide effective traffic scheduling while attempting to mitigate the aforementioned problems depending on certain estimates of path and connection quality [13]–[15].

1.2. Contributions and Highlights of this Paper. To improve multi-path data transmission performance by utilizing delay variations, this research suggests an approach technique to MP-TCP. When scheduling data over numerous paths, our approach effectively accounts for varied path characteristics and handles unjustified cwnd growth modifications. We have taken into consideration delay variation as a statistic that, in the end, shows the capacity of the resources for every path. In addition, we have proposed an adaptive rapid retransmission policy based on delay variation that further controls the suitable transmission rate for every path. As a matter of fact, our approach effectively predicts congestion during the rising phase by periodically evaluating the increasing delay variation of the data segments that have been transferred over a path. Then, our approach modifies the cwnd growth scheme for the congested path.

- We propose policy that is adaptive and flexible multi-path data transfer policy for MP-TCP.
- Also proposes delay variations based adaptive data packet scheduling and adaptive fast retransmission policy to MP-TCP.
- We confirmation of the effectiveness of the proposed approach by conducting extensive experiments and by doing comprehensive evaluation. The validation of our approach has been carried out on ns-2.
- Our offers Improvement of 46 percent, 19 percent, and 34 percent better in throughput performance for varied packet loss rate, BW, and path delay cases respectively. Additionally, the approach offers improved file transfer time as compared to MP-TCP.

2. Related Work. The ultimate goal of this section is to outline for the reader the main concerns affecting various SCTP and MP-TCP based approaches. Additionally, in this area, new concerns related to these difficulties have been included. In addition, we will comprehend in this section how study endeavors have been undertaken to address all these problems. Additionally, even if after resolving all of these difficulties, what new issues have surfaced? There are two types of multi-path Layer-4 protocols: connection-oriented and connectionless (also known as multi-path real-time protocols). Still, this section will focus exclusively on connection-oriented protocols.

When TCP was first designed, end hosts could only connect using single interface support that is, with the same IP address. However, the increased usage of complex networking devices has led to the creation of end-hosts that can handle numerous interfaces. In the modern era of communication, cellphone networks and Wi-Fi are extensively accessible in urban areas. It would be desirable for users of smartphones and tablets to be able to start a connection, especially a TCP connection in a Wi-Fi hotspot and continue it later to their 3G interface [16] [17].

As a result, extremely intelligent middle-boxes are developed, thereby dispelling the need for ETE design. Moreover, these sophisticated middle-boxes have provided a host of benefits, including speed enhancements, security (firewalls), usage (load balancers), etc. But the rapid collapse of these clever middle-boxes also resulted in increased complexity when it came to Layer-4's design protocols. However, SCTP's adoption is quite challenging in the current Internet architecture since it relies on the very simple assumption that the received data segments remain identical. As of now, MP-TCP is the only protocol that was created with middle-box functionality in mind. However, under a few unique circumstances, MP-TCP's performance might be considerably impacted when working with certain middle-boxes as well. Consequently, a large amount of study is needed to handle middle-boxes efficiently [3] [18].

2.1. Multiple path based Proposals. Researchers were certain that multi-pathing was inevitable because new designs are always striving to develop more and more workable solutions. This in turn led to the development of MP-TCP. Originally, several congestion control strategies that immediately extended TCP New Reno for the aim of creating the MP-TCP policy, also known as the Linked Increase Algorithm (LIA) [19], i.e., recommended policies, directly initiate the TCP New Reno's capability on each sub-path independently. When the available network paths share bottleneck links with the network paths used by MP-TCP users the network's single-path TCP users may experience significant unfairness as a result of this direct extended version. To structure an effective multi-path Layer-4 protocol that is specifically compatible with the conventional TCP (i.e., Coupled Congestion Control (CCC) algorithm), several academics have proposed several techniques [20]. These suggested methods in [21]–[24] are most suited for situations where similar or minimal differences in Round Trip Times (RTTs) have been recorded, as they typically only use the best paths that are available to the users. However, these algorithms experience floppiness and lower responsiveness.

First of all, these algorithms are not always able to adjust quickly. Specifically, they are not able to explore the paths that have a greater channel and congestion-induced loss probability, which results in a significantly lower level of responsiveness. Second, the network exhibits extreme floppiness according to these techniques. Peng et al. [25] suggested that the half-coupled congestion control approach is inhospitable to single-path TCP users. It inflates during different RTT iterations of every feasible sub-path.

The authors have subsequently defined design standards that provide assurance of existence, stability, and uniqueness of the system. Their method, known as the Balanced Link Adaptation Algorithm (BALIA) [26], was primarily concerned with performance measures including cwnd fluctuations, TCP friendliness, and receptiveness. Additionally, Lubna et al. [27] proposed the Dynamic OLIA (DOLIA) policy, which successfully restrained MP-TCP's aggression about the cwnd expansion. Many adaptive scheduling policies that guarantee in-order data delivery have already been proposed to mitigate the buffer-blocking issue. In the meantime, researchers have previously done a great deal of research on topics like path heterogeneity [28]–[30], congestion control techniques [31]-[34], enhanced traffic management schedulers [21][24][35].

The MP-TCP SBD, or Shared Bottleneck Detection based Coupled Congestion Control [36] is an issue to be put light on. The performance of other sub-flows is impacted by the extra loss and delayed sub-flow, as demonstrated by the writers. As a result, the throughput performance is eventually hampered and the bottleneck of many MP-TCP connections develop. The authors recommended using fountain codes to effectively manage the variable attributes of numerous pathways to mitigate these impacts.

For multipath transmission, Thomas et al. [37] proposed a hybrid congestion control policy. The Multi-flow Congestion Control with Network Assistance (MFCCNA) the policy has been provided by the authors; their approach dynamically examines the available topological information about the network. The ultimate goal of MFCCNA is to improve network resource consumption without sacrificing friendliness. Nevertheless, the harshness of the window growth program was not taken into account by MFCCNA thus causes a decline in performance about packet losses. Until its cwnd is entirely filled with data, the current scheduler (default), minimal RTT (RTTmin) first, distributes data packets to the fastest available sub-flow. The leftover packets are then suitably assigned to other available sub-flows. Although this method outperforms the traditional round robin data scheduling scheme in the majority of circumstances, new research indicates that the RTTmin approach still needs more thorough assessment and consideration when the available network pathways have different attributes.

Thus, this / it further contributes to the Head-of-Line (HoL) [38] [39] blocking problem, which raises the likelihood of receiving data packets in a disordered manner at the receiver buffer and, in turn, lowers the throughput performance at the application level. As a result, many improved dynamic scheduling techniques have been created in the past for efficiently dividing up data packets among several open network channels. While the most recent delay-aware scheduling policies can enhance MP-TCP performance, they are unable to provide reliability assurances. However, because of the diverse environment, delay aware scheduling strategies need to handle the problem of long retransmissions and highly unreliable features of multi-path sessions.

2.2. Security, Reliability and compatibility of MPTCP. Firstly security, because multipath routing fragments cross-path data, firewalls and virus scanners become ineffective when they can only able to track traffic on one path. While MPTCP doesn't enforce encryption on its own, it can be used in conjunction with secure encryption protocols such as TLS (Transport Layer Security) or IPsec (Internet Protocol Security). When data is encrypted using these protocols, it remains confidential and secure against eavesdropping attacks, ensuring that data transmitted over MPTCP connections is protected [40].

Secondly reliability, MPTCP enhances reliability by providing fault tolerance, increasing throughput, adapting to network fluctuations, supporting seamless handover, implementing selective retransmission, and employing efficient congestion control mechanisms.

Thirdly for compatibility with existing protocols, a thorough study is being done before implementing MPTCP. The middleboxes play a very important role in the performance of the protocol. Outside protocols see MPTCP is an inside part of TCP only. Therefore our introduced changes are also not affecting other protocols that deal with the transmission of data in the network.

2.3. The Proposed Method. When evaluating the wired network scenario during multi-homed communication, there is typically a chance that the available channel parameters (i.e., path latency, BW, and loss rate) will vary greatly. Therefore, if MPTCP uses the entire available network interfaces for multi-path transmissions, it will surely result in the problem of packet reordering in the framework. As a result, this strategy severely degrades the network's throughput performance. In order to guarantee efficient support for MP-TCP, this article introduces the assorted path scheduling approach. Abrupt traffic management (i.e., scheduling) schemes and ridiculously incorrect congestion window growth adjustments are guaranteed to be efficiently handled by our proposed policy. To achieve this, our policy takes into account the quality of any network path that is participating in data transmission and also modifies the data scheduling approach.

The scalability of the proposed work is wherever MPTCP is being used; the proposed approach can be used i.e. on more complex and large networks too. The proposed approach takes RTT variations and congestion window capacity as its metric to decide where to schedule next to. Calculating RTT each time and scheduling desired traffic is a bit complex procedure but ns-2 does this at ease. Also, while scaling Multipath TCP to larger networks or more complex environments present challenges related to overhead, path management, resource constraints, interoperability, security, privacy, and management, these challenges can be addressed through optimizations in MPTCP implementations, efficient path management strategies, resource optimization techniques, interoperability efforts, enhanced security measures, and effective management and monitoring tools. By addressing these considerations, MPTCP can be effectively scaled to meet the demands of diverse networking environments, providing improved performance, reliability, and flexibility for data transmission over multipath networks.

Delay variations, an essential metric that effectively illustrates the representation of resource availability for a candidate path, is what our policy has employed for this. In addition, an adaptive rapid retransmission policy based on the delay variations metric is presented in this study for the same. Our approach bears the obligation to control an effective transmission management plan in line with the available network pathways varying delay.

$$\text{Packets\_sent(on\_a \_ path)} = \text{cwnd\_currentpath -unacknowledged\_packets} \tag{2.1}$$

$$\text{Waiting\_ time} = \text{immediately\_ sent\_ packets - packets\_ not\_sent} \tag{2.2}$$

3. Problem Identification. Indeed, a packet scheduler is required to enable the use of various accessible pathways in MP-TCP. Selecting which data packet should be sent via which open path the primary duty of the packet scheduler is. On the other hand, MP-TCP has to transmit packets at the receiver's end to guarantee system reliability in the same order in which they are sent from the sender's end. Specifically, the features of the dissimilar pathways, such as packet loss rate, delay, and BW availability, will consistently result in data being received in an erratic manner at the recipient's end. Nonetheless, the receiver buffer included in the standardized MP-TCP eventually helps the protocol reorder data packets originating from all of the accessible interfaces. However, the MP-TCP receiver buffer capacity is extremely constrained; as a result, the buffer becomes blocked when the frequency of receiving increases and the amount of unordered data packets increases. The MP-TCP sender maybe forced by this circumstance to send the new data packets throughout the stable network pathways that are open for transmission. The reason for this is that there are insufficient buffer spaces at the receiver's end for the incoming fresh data packets. This indicates that unstable paths (i.e., paths with greater delay variability) equally, disrupt the available stable routes inside the MP-TCP connection for the delivery of data packets. Consequently, the MP-TCP application-level throughput performance may be significantly decreased by unstable pathways utilized for multi-path data transfer.

3.1. Proposed Solution. The suggested approach's primary goal is to estimate unstable pathways promptly. As a result, our approach can effectively handle irrational congestion window growth adjustments in addition to scheduling transmissions via alternative stable pathways. More specifically, our suggested method arranges the transmission load over various available stable sub-paths suitably and estimates the paths with higher delay fluctuations. According to Equation 3, the propagation delay (Prop-Delay), processing delay (Proc-Delay), queueing delay (Queue-Delay), and transmission delay (Trans-Delay) generally make up the path delay (Path-Delay).

We quickly discuss the aforementioned network delays that ultimately led to the overall packet delay to illustrate the impact of path delay. Trans-Delay: - The amount of time required by the sender's Layer-1, or Physical Layer, to transfer every bit onto the wire, Prop-Delay: - The flight time of bits over the connected channel is how it is defined. Queue-Delay: - A packet must wait an endless period of time in a router's queue before it can be processed; this time is dependent on the router's traffic load. The precise information about the level of network congestion is provided by Queue-Delay. Proc-Delay: - A packet experiences a continuous delay at both the sender's and the recipient's end.

$$PATH\_Delay(i) = PRO\_Delay(i) + TRANS\_Delay(i) + QUEUE\_Delay(i) + PROC\_Delay(i) \quad (3.1)$$

This delay may be caused by the time it takes for analog data to be converted to digital form at the sender's end and then packetized through multiple layers of operating protocols before the data are sent to the Physical Layer for transmission.

Similar to this, the behavior of the underlying hardware, the Operating System (OS), and the kind of application data being considered may have an impact on this delay at the receiver level. However, we use path delay in our proposed approach to explicitly evaluate the delay fluctuations. Therefore, it is not necessary to evaluate the aforementioned delay kinds on the proposed policy in detail.

3.2. Validity of Proposed Work. This section explains the validation of the proposed policy to mitigate losses and increase bandwidth utilization. Our proposed policy anticipates congestion in the network and timely assessment of dealing with that occurrence of congestion. The assessment is done by checking the variations in RTT of each path; if some sudden changes are there in RTT, it means congestion may occur. So this is the Congestion Avoidance phase. Our approach suggests an improvement in the CA phase only so that the utilization of the bandwidth is done and overall performance of the network is being improved. Therefore it is easily implementable for other researchers also. The Proposed algorithm is also there to help understand the proposed approach and simulated results.

3.3. Generalizability and scaling Multipath TCP (MPTCP). MPTCP have broad applicability across various network configurations, traffic patterns, and application scenarios. Here's an exploration of the generalizability of these findings: By addressing scalability challenges and optimizing MPTCP deployments for diverse use cases, organizations can leverage the benefits of multipath communication to enhance performance, reliability, and flexibility in their networks and applications.

1. *Network Configurations:* MPTCP's scalability solutions apply to different network configurations, including local area networks (LANs), wide area networks (WANs), data center networks, and wireless networks. Regardless of the specific topology or technology used in the network, MPTCP's ability to leverage multiple paths simultaneously can improve performance and reliability by utilizing available resources efficiently and mitigating congestion.
2. *Traffic Patterns:* MPTCP's scalability considerations are relevant for diverse traffic patterns, including bursty traffic, streaming media, interactive applications, and large-scale data transfers. By dynamically adapting to changes in traffic conditions and network capacity, MPTCP can optimize data transmission to meet the requirements of different applications and traffic patterns, ensuring reliable communication and efficient resource utilization.
3. *Application Scenarios:* MPTCP's scalability solutions are applicable to various application scenarios, such as cloud computing, content delivery, real-time communication, and mobile networking. Whether it's distributing workload across multiple servers in a cloud environment, delivering content to geographically distributed users, maintaining low-latency communication in real-time applications, or supporting seamless mobility for mobile devices, MPTCP's multipath capabilities can enhance performance, reliability, and flexibility across a wide range of use cases.
4. *Heterogeneous Environments:* MPTCP's scalability findings are relevant for heterogeneous environments comprising different network technologies, devices, and operating systems. Whether it's integrating MPTCP support into legacy systems, ensuring interoperability between diverse network components, or accommodating heterogeneous network conditions, the proposed solutions can help overcome compatibility challenges and promote the adoption of MPTCP across heterogeneous environments.
5. *Future Trends:* As networking technologies continue to evolve, including the proliferation of Internet of Things (IoT) devices, the deployment of 5G networks, and the emergence of edge computing, MPTCP's scalability solutions will remain relevant for addressing the scalability, performance, and reliability requirements of future network architectures and applications. By adapting to emerging trends and technologies, MPTCP can continue to provide value in an ever-changing networking landscape.

3.4. Proposed Policy. Each TCP sub-flow in MP-TCP maintains an exclusive congestion window. Additionally, if three duplicate ACKs (dupACKs) are received, the congestion window is reduced automatically by half. On the other hand for MPTCP, coupled congestion control algorithm (CCCA) is used to regulate the augment of the congestion window of every sub-flow at the MP-TCP flow level. CCCA also does resource utilization at each available TCP subflows and try to deliver fairness among them. The congestion window growth of each path is different and availability of each path leads to delay variation in congestion windows which estimates false congestion sometimes. This difference in RTT of available path is one of the reasons why the performance of MPTCP degrades. The goal of this approach is to combine the higher BW of several accessible links while avoiding or avoiding the more aggressive MP-TCP protocol than regular TCP flows on each link that is used. This is done to prevent MP-TCP from using excessive amounts of resources and ensure TCP friendliness. Moreover, packet loss caused by duplicate ACKs is the only factor in this technique that lowers the congestion window. Consequently, there may be a significant variation in the congestion window of each path. As a result, there maybe a significant path delay difference or variation, which is detrimental to the throughput performance at the application level.

In our approach, the capacity and delay of each available path is being monitored continuously which gives the actual state of each path so that the transmission of data is done accordingly. In other words, we can get the picture of stable and unstable paths for further transmission of data. Growth of each congestion window is done and explained with the help of equation 3.2 -3.3. Hence in our approach, the paths have been identified with actual window capacity and despite following traditional RR policy; data transmission is done according to delay and capacity of each path. Hence data is transmitted fast and efficiently. More precisely, the Slow-Start (SS) of MP-TCP is designed to detect the bandwidth availability of the path in order to prevent network congestion. The window growth mechanism moves from the SS phase to the Congestion Avoidance (CA) phase as soon as the congestion window growth reaches a certain extent (SSThresh). The sender then transitions to the CA phase by modifying its SSThresh as soon as feasible, as detected by the timeout mechanism, should a packet loss occur. When in slow start, the window increases after receiving an acknowledgement each time as follows:

$$\min((\alpha/cwnd\_total\_paths), 1/cwnd\_current\_path) \tag{3.2}$$

where α can be calculated as:

$$\alpha = \frac{cwnd\_total\_paths(\max(cwnd\_current\_path/rtt\_current\_path^2)}{\sum(cwnd\_current\_path/rtt\_current\_path)} \tag{3.3}$$

Eq. 3.2 and Eq. 3.3 can be simplified when substituting value of alpha from Eq. 3.2.1 to Eq. 3.2 and is given as follows:

$$\alpha = cwnd\_total\_paths \cdot (\max(cwnd\_i/(cwnd\_current\_path)^2) \tag{3.4}$$

The evaluation of the suggested policy's cwnd growth policy has been conducted in accordance with the existing delay variations, which equal a path. The optimal congestion window growth behavior has been found, fully using the likely channel use. It guides performance improvement in relation to bandwidth and total number of network timeouts. Let the pathways be Pi = (P1, P2, P3 ... Pn). Di = (D1, D2, D3 ... Dn) indicates the RTT latency of every path − its recurrence. The first measured RTT must be used to initialize the minimum RTT, or ”RTTMin.”

3.5. Algorithm. See Algorithm 1.

4. Assessment of performance. This section conducts extensive experiments in Network Simulator-2 (ns-2) to validate our proposed technique and evaluates its effectiveness against current multi-path techniques.

4.1. Experimental design and simulation setup. The simulation has been run entirely on NS-2.35.

About real world testing: for implementing MPTCP we need more than two systems having configuration that supports our protocol. Out of which windows still don't support MPTCP protocol so Linux and IOS

Algorithm 1 An algorithm for scheduling of Data

```
Requirement: RTT_curr(i), RTT_min_delay($i$), $\alpha$Thresh, ssthresh(Segment)
Begin: Calculate initial ssthresh in terms of Bytes
ssthresh$_i$(Bytes) $\leftarrow$ ssthresh(Segment) * Maximum Segment Size (MSS)
For every SACK received (at the sender side for each destination):
Calculate delay
IF cwnd$_i$ <ssthresh$_i$
    cwnd$_{i+1}$ $\leftarrow$cwnd$_i$+1 // (SSphase)
        ELSEIF $\mu_i > \alpha$Thresh
            cwnd$_{i+1}$ $\leftarrow$ cwnd$_i$
        ELSE
            path_order$_i$ $\leftarrow$cwnd$_i$
              RTT_curr($i$) $\leftarrow$RTT_min($i$) // (Congestion Avoidance Phase)
```

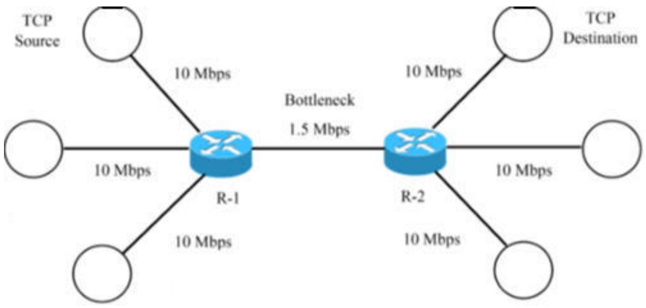


Fig. 4.1: Simulation topology

should be configured in the system. The simulator we used to perform test cases is ns-2.34 which is widely accepted and known for running almost every networking protocol.

The network topology employed in the simulation is displayed in Figure 4.1. One source is used in the topology with one destination with two network interfaces. Figure 4.1 displays the bandwidth and delay of each link with background traffic source and destination. One TCP source is used as a background traffic generator to check the performance of the proposed approach while traffic is there. Each of the nodes are connected to Router 1 with 10 Mbps bandwidth and with Router 2 also same 10Mbps bandwidth while both of the routers R1 and R2 are connected and forming a bottleneck situation with 1.5 Mbps bandwidth. The sender side has background traffic attached to it as TCP Source and UDP Source creating a bottleneck at the middle of the topology with routers R-1 and R-2.

4.2. Simulation and Results. Results of the work are being compared and presented graphically in this section. We have implemented different network configurations, and traffic patterns. The clear comparison of results is done on the topology being mentioned in Figure 4.1. As far as traffic patterns we have used TCP for background traffic we have attached another figure showing throughput in case of UDP as background traffic now. Only TCP and UDP can be used as background traffic for MPTCP.

The change in throughput (Kbps) during the course of the simulation is displayed in Figure 4.2 and 4.3. When a receiver buffer is employed, the MP-TCP and suggested policy results show the traffic across PATH-1 and PATH-2, respectively. Because both MP-TCP and the proposed policy promptly investigate the available channel capacity, their throughput initially climbs significantly. The suggested method transfers more data on a path with lower latency and higher bandwidth by using the path capacity and latency as a consideration in data chunk scheduling.

Additionally, the Slow Start algorithm repeatedly doubles the size of the cwnd. In particular, packet loss and recovery cause a rapid change in throughput by its half. As a result, both policies recover from

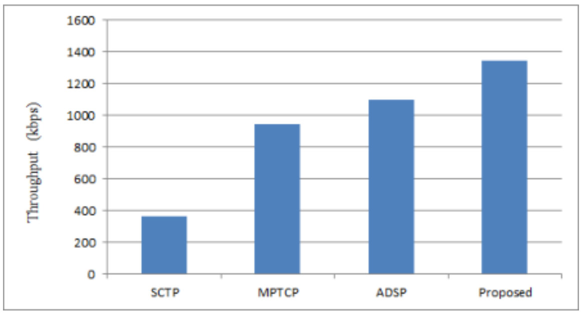


Fig. 4.2: Throughput comparison of SCTP, MPTCP and Proposed Approach

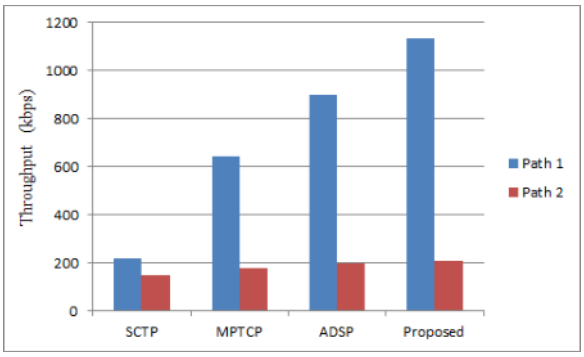


Fig. 4.3: Throughput comparison of SCTP, MPTCP and Proposed Approach

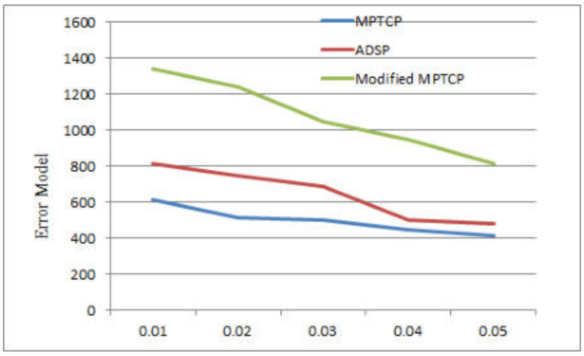


Fig. 4.4: Packet Loss Ratio of MPTCP and our Proposed Approach

losses following appropriate cwnd modifications and quick retransmissions. As a result, the proposed one uses adaptive rapid retransmission policy, RTT variations, and strong packet loss and unordered delivery mitigation to provide higher throughput performance.

The fluctuations in average throughput (Kbps) during the simulation period are displayed in Figure 4.4. The purpose of this experiment is to confirm that every simulated strategy for effectively managing packet losses, which significantly affect average throughput. The average throughput for all the simulated techniques continuously drops as the PLR grows, as illustrated in Figure 4.4. The retransmission timeout is then examined when path-1 packet loss rate is 1 percent and path-2 packet loss rate ranges from 1 percent to 10 percent in

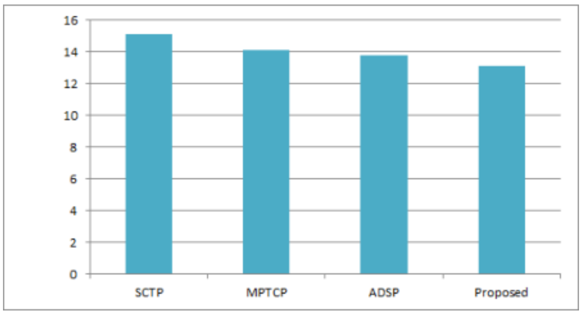


Fig. 4.5: Retransmission time in MPTCP and our Proposed Approach

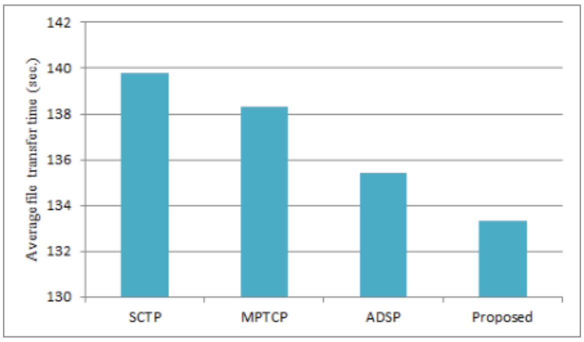


Fig. 4.6: Average file transfer time in SCTP, MPTCP and Proposed Approach

Figure 4.5. The throughput is directly impacted by the retransmission timeout. Out of all the methods utilized for comparison, the suggested approach has the lowest timeout. Consequently, the suggested approach has a greater average usage than SCTP and MPTCP as shown in Figure 4.5. Different multipath protocols have different file transfer times are displayed in Fig. 4.6, with file sizes ranging from 10MB to 90MB. It indicates that the file transfer time grows together with the file size. The average file transfer time improvement using the suggested method is 4.6 percent when compared to MPTCP and SCTP. Therefore, compared to the previously discussed methods, the suggested solution performs better overall in a symmetric packet loss situation.

The suggested technique transmits data over several paths while optimizing network use by using bandwidth and delay-aware scheduling. In contrast to SCTP and MPTCP, the suggested technique transmits files of any size faster. In Figure 4.7, it is illustrated that file transfer times are different for files ranging in size from 10MB to 90MB therefore it demonstrates that file transfer times rise in tandem with file sizes. The data pieces are being distributed equally across several pathways by both strategies. The suggested approach factors in route latency and bandwidth while arranging data chunks. Thus, a lot of data is scheduled on the high bandwidth and minimum delay line by the suggested strategy.

5. Conclusion. This study introduced an approach that is bendable and adaptive multi-path data transmission scheme, in response to the constraints posed by various path characteristics in concurrent multi-path transmission design. MP-TCP transfer scheme. Rather than arbitrarily allocating traffic across several viable network channels without taking into account the differences in path characteristics, the proposed one considers each path's quality and adjusts the scheduling criteria accordingly. The proposed policy handles the buffer-blocking issue at the receiver side and the sudden congestion window growth adjustments issue at the transmitter side with its RTT variation-based adaptive scheduling and rapid retransmission technique. Using the idea of evaluating RTT variations to determine the traffic state of a path, the RTT variation-based adaptive scheduling technique operates. Additionally, the adaptive rapid retransmission technique based on RTT variations controls the volume of traffic (transmission rate) in relation to RTT variations. The simulation results

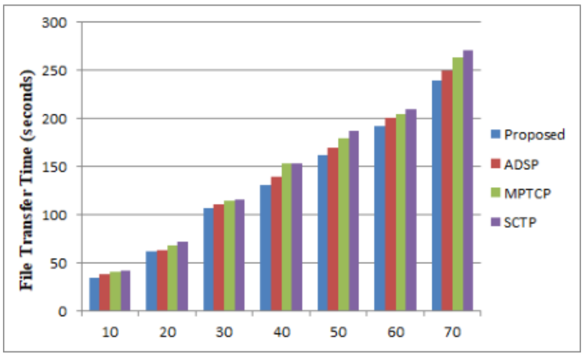


Fig. 4.7: File transfer time in SCTP, MPTCP and Proposed Approach

show how effective our proposed approach is in terms performance and average throughput. The proposed approach gives 56 percent, 19 percent, and 36 percent greater throughput than MP-TCP, respectively and the proposed approach provides 50 percent better FTT.

REFERENCES

[1] J. Postel, Transmission Control Protocol, Darpa Internet Program, September 1981. RFC 793. [Online]. Retrieved: https://datatracker.ietf.org/doc/html/rfc793.

[2] J. Postel, User Datagram Protocol, RFC768, The Internet Engineering Task Force, California, USA, 2007. RFC 768. [Online]. Retrieved: https://datatracker.ietf.org/doc
/html/rfc768.

[3] Habib, S., Qadir, J., Ali, A., Habib, D., Li, M., and Sathiaseelan, A. (2016). The past, present, and future of transport-layer multipath. Journal of Network and Computer Applications 75, 236–258.

[4] Khalili, R., Gast, N., Popovic, M., and Le Boudec, J.Y. (2013). MPTCP is not pareto-optimal: Performance issues and a possible solution. IEEE/ACM Transactions on Networking 21, 1651–1665.

[5] Raiciu, C., Paasch, C., Barre, S., Ford, A., Honda, M., Duchene, F., Bonaventure, O., and Handley, M. (2012). How hard can it be? Designing and implementing a deployable multipath TCP. In Proceedings of NSDI 2012: 9th USENIX Symposium on Networked Systems Design and Implementation, (USENIX Association), pp. 399–412.

[6] Honda, M., Nishida, Y., Eggert, L., Sarolahti, P., and Tokuda, H. (2009). Multipath Congestion Control for Shared Bottleneck. International Workshop on Protocols for Future, Large-Scale and Diverse Network Transports (PFLDNeT).

[7] Paasch, C., and Enture, O.B. (2014). Multipath TCP. Communications of the ACM 57, 51–57.

[8] Verma, L.P., Sharma, V.K., and Kumar, M. (2018). New delay-based fast retransmission policy for CMT-SCTP. International Journal of Intelligent Systems and Applications 10, 59–66.

[9] Shailendra, S., Bhattacharjee, R., and Bose, S.K. (2011). MPSCTP: A simple and efficient multipath algorithm for SCTP. IEEE Communications Letters 15, 1139–1141.

[10] Xu, C., Zhao, J., and Muntean, G.M. (2016). Congestion Control Design for Multipath Transport Protocols: A Survey. IEEE Communications Surveys and Tutorials 18, 2948–2969.

[11] Ferlin, S., Dreibholz, T., and Alay, Ö. (2014). Multi-path transport over heterogeneous wireless networks: Does it really pay off? In 2014 IEEE Global Communications Conference, GLOBECOM 2014, (Institute of Electrical and Electronics Engineers Inc.), pp. 4807–4813.

[12] Ha, S., Le, L., Rhee, I., and Xu, L. (2007). Impact of background traffic on performance of high-speed TCP variant protocols. Computer Networks 51, 1748–1762.

[13] Noura, H.N., Melki, R., and Chehab, A. (2022). Network coding and MPTCP: Enhancing security and performance in an SDN environment. Journal of Information Security and Applications 66.

[14] Ha, T., Masood, A., Na, W., and Cho, S. (2023). Intelligent Multi-Path TCP Congestion Control for video streaming in Internet of Deep Space Things communication. ICT Express 9, 860–868.

[15] Li, M., Lukyanenko, A., Tarkoma, S., Cui, Y., and Ylä-Jääski, A. (2013). Tolerating path heterogeneity in multipath TCP with bounded receive buffers. In Performance Evaluation Review, pp. 375–376.

[16] Lin, J., Cui, L., Zhang, Y., Tso, F.P., and Guan, Q. (2019). Extensive evaluation on the performance and behaviour of TCP congestion control protocols under varied network scenarios. Computer Networks 163.

[17] Wang, Z., Zeng, X., Liu, X., Xu, M., Wen, Y., and Chen, L. (2016). TCP congestion control algorithm for heterogeneous Internet. Journal of Network and Computer Applications 68, 56–64.

[18] Bonaventure, O., Handley, M., and Raiciu, C. (2012). An Overview of Multipath TCP. USENIX Login 17–23.

[19] Lubna, T., Mahmud, I., and Cho, Y.Z. (2020). D-LIA: Dynamic congestion control algorithm for MPTCP. ICT Express 6,

263–268.

[20] Verma, L.P., Sharma, V.K., Kumar, M., and Mahanti, A. (2022). An adaptive multi-path data transfer approach for MP-TCP. Wireless Networks 28, 2185–2212.

[21] Ferlin, S., Alay, O., Mehani, O., and Boreli, R. (2016). BLEST: Blocking estimation-based MPTCP scheduler for heterogeneous networks. In 2016 IFIP Networking Conference (IFIP Networking) and Workshops, IFIP Networking 2016, (Institute of Electrical and Electronics Engineers Inc.), pp. 431–439.

[22] Dong, P., Yang, W., Tang, W., Huang, J., Wang, H., Pan, Y., and Wang, J. (2018). Reducing transport latency for short flows with multipath TCP. Journal of Network and Computer Applications 108, 20–36.

[23] Chen, D., Gao, D., Jin, L., Quan, W., and Zhang, H. (2023). ADAS: Adaptive Delay-Aligned Scheduling for Multipath Transmission in Heterogeneous Wireless Networks. Peer-to-Peer Networking and Applications 16, 1583–1595.

[24] Lim, Y.S., Towsley, D., Nahum, E.M., and Gibbens, R.J. (2017). ECF: An MPTCP path scheduler to manage heterogeneous paths. In CoNEXT 2017 - Proceedings of the 2017 13th International Conference on Emerging Networking EXperiments and Technologies, (Association for Computing Machinery, Inc), pp. 147–159.

[25] Peng, Q., Walid, A., Hwang, J., and Low, S.H. (2016). Multipath TCP: Analysis, Design, and Implementation. IEEE/ACM Transactions on Networking 24, 596–609.

[26] Kato, T., Diwakar, A., Yamamoto, R., Ohzahata, S., and Suzuki, N. (2019). Experimental analysis of MPTCp congestion control algorithms; Lia, olia and Balia. In Multi Conference on Computer Science and Information Systems, MCCSIS 2019 - Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019, (IADIS Press), pp. 135–142.

[27] Lubna, T., Mahmud, I., Kim, G.H., and Cho, Y.Z. (2021). D-olia: A hybrid mptcp congestion control algorithm with network delay estimation. Sensors 21.

[28] Li, M., Lukyanenko, A., Tarkoma, S., Cui, Y., and Ylä-Jääski, A. (2013). Tolerating path heterogeneity in multipath TCP with bounded receive buffers. In Performance Evaluation Review, pp. 375–376.

[29] Vo, P.L., Le, T.A., Lee, S., Hong, C.S., Kim, B., and Song, H. (2014). Multi-path utility maximization and multi-path TCP design. Journal of Parallel and Distributed Computing 74, 1848–1857.

[30] Ferlin, S., Dreibholz, T., and Alay, Ö. (2014). Multi-path transport over heterogeneous wireless networks: Does it really pay off? In 2014 IEEE Global Communications Conference, GLOBECOM 2014, (Institute of Electrical and Electronics Engineers Inc.), pp. 4807–4813.

[31] Zhou, D., Song, W., and Cheng, Y. (2013). A study of fair bandwidth sharing with AIMD-based multipath congestion control. IEEE Wireless Communications Letters 2, 299–302.

[32] Carofiglio, G., Gallo, M., and Muscariello, L. (2016). Optimal multipath congestion control and request forwarding in information-centric networks: Protocol design and experimentation. Computer Networks 110, 104–117.

[33] Tao, Y., and Huang, P. (2014). MPTCP congestion control algorithm based on the fairness of bottleneck. In Applied Mechanics and Materials, (Trans Tech Publications), pp. 3995–4000.

[34] Mahmud, I., Lubna, T., Song, Y.J., and Cho, Y.Z. (2020). Coupled multipath BBR (c-MPBBR): An efficient congestion control algorithm for multipath TCP. IEEE Access 8, 165497–165511.

[35] Yang, W., Dong, P., Tang, W., Lou, X., Zhou, H., Gao, K., and Wang, H. (2018). A MPTCP scheduler for web transfer. Computers, Materials and Continua 57, 205–222.

[36] Wei, W., Xue, K., Han, J., Wei, D.S.L., and Hong, P. (2020). Shared Bottleneck-Based Congestion Control and Packet Scheduling for Multipath TCP. IEEE/ACM Transactions on Networking 28, 653–666.

[37] Thomas, Y., Xylomenos, G., and Polyzos, G.C. (2020). Multipath congestion control with network assistance. Computer Communications 153, 264–278.

[38] Ha, T., Masood, A., Na, W., and Cho, S. (2023). Intelligent Multi-Path TCP Congestion Control for video streaming in Internet of Deep Space Things communication. ICT Express 9, 860–868.

[39] Choi, K.W., Cho, Y.S., Aneta, Lee, J.W., Cho, S.M., and Choi, J. (2017). Optimal load balancing scheduler for MPTCP-based bandwidth aggregation in heterogeneous wireless environments. Computer Communications 112, 116–130.

[40] Popat, K., Kapadia, V.V. (2021). Multipath TCP Security Issues, Challenges and Solutions. In: Bhattacharya, M., Kharb, L., Chahal, D. (eds) Information, Communication and Computing Technology. ICICCT 2021. Communications in Computer and Information Science, vol 1417. Springer, Cham. https://doi.org/10.1007/978-3-030-88378-22.

[41] Zhao Baosen, Yang Wanghong, Du Wenji, Ren Yongmao, Sun Jianan, Wu Qinghua, Zhou Xu. (2024). A multipath scheduler based on cross-layer information for low-delay applications in 5G edge networks. Computer Networks vol 224. Elsevier. https://https://doi.org/10.1016/j.comnet.2024.110333.

[42] Amend, M., and Rakocevic, V. (2024). Cost-efficient multipath scheduling of video-on-demand traffic for the 5G ATSSS splitting function. Computer Networks 242.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Mar 13, 2024
Accepted: Jun 7, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 349–360, DOI 10.12694/scpe.v26i1.3525

MEMORY, CHANNEL AND PROCESS UTILIZATION FOR FUZZY BASED CONGESTION DETECTION AND AVOIDANCE SCHEME IN FLYING AD HOC AND IOT NETWORK

MAHENDRA SAHARE [*]AND PRITI MAHESHWARY [†]

Abstract. UAVs are flying in the air at different speeds and continuously forwarding the collected information to other UAVs or IoT devices in FANET. UAVs are playing an important role in data collection from places where humans can't reach them easily. The UAVs are intelligent devices, and these devices have sufficient bandwidth and memory for data forwarding and storing. The role of UAVs is specific, and they have the reflexibility to change the battery and control the data interval to control the congestion in network. The IoT devices with FANET can transfer the valuable data to other IoT devices for verification and matching. The proper utilization of bandwidth, memory, energy and processing capability are able to increase the Quality of Service (QoS) in FANET. In this paper, proposed the Memory, channel and Process utilization for Fuzzy based (MCPFB) for congestion detection and avoidance scheme to improve bandwidth utilization, energy consumption in FANET with the IoT network. primarily aims to identify and prevent network congestion, which is crucial for maintaining the QoS requirements and ensuring reliable communication. Congestion is a phenomenon that arises when the volume of data transmitted across a network exceeds its capacity. These factors can lead to disruptions, reduced efficiency, and potential data loss in communication networks such as Flying Ad Hoc Networks (FANETs). To effectively handle congestion in FANET and provide reliable communication in challenging and dynamic environments, it is crucial to employ efficient resource management, intelligent algorithms, and adaptable protocols. The process of designing fuzzy rules for Flying Ad Hoc Networks (FANET) entails developing a set of guidelines that utilize fuzzy logic to make decisions pertaining to different parts of the network. The MCPFB is better than the previous BARS approach in terms of different performance metrics.

Key words: Bandwidth, Congestion, Energy, MCPFB, IoT, FANET

1. Introduction. In any network communication between devices play an important role in exchanging data from one place to another, but it depends on various factors i.e. communication medium, channel availability, intermediate devices, queue capacity, processing capacity of network devices, etc [1] [2]. In the new era of communication technology wireless communication plays a vital role in providing communication anywhere at any time which is further categorized in three ways FANET, MANET, and VANET. This paper works under FANET, Flying Ad Hoc Network (FANET) is a specialized kind of mobile ad hoc network (MANET) that facilitates communication between unmanned aerial vehicles (UAVs) or drones [1] [2]. Flying ad hoc network completely depends on intermediate nodes which help to provide a route from one device to another using FANET routing protocol, but the data exchange from device to device is not an easy task it requires sufficient channel bandwidth, processing power, node energy, node mobility, topology monitoring, etc. due to all this requirement, it a chance to challenge of network congestion. In this paper, our main focus is to detect and avoid network congestion which is further useful to maintain the network quality of service requirement and provide reliable communication.

In the given figure 1.1 number if UAVs are five and only two IoT devices are collecting the information from UAVs. UAVs are also connected to Base station. When the amount of data sent via a network surpasses its capacity, a phenomenon known as congestion occurs [3]. This can cause delays, decreased performance, and even packet loss in communication networks like Flying Ad Hoc Networks (FANETs). Efficient resource management, smart algorithms, and adaptable protocols are needed to manage FANET congestion and guarantee dependable communication in difficult and ever-changing conditions. In this paper detect and avoid congestion using fuzzy rules, designing fuzzy rules for Flying Ad Hoc Networks (FANET) involves creating a set of guidelines based on

[*]Dept. of Computer Science & Engineering, Rabindranath Tagore University Bhopal (MP), India (mahendrasahare1110@gmail.com)

[†]Dept. of Computer Science & Engineering, Rabindranath Tagore University Bhopal (MP), India (pritimaheshwary@gmail.com)

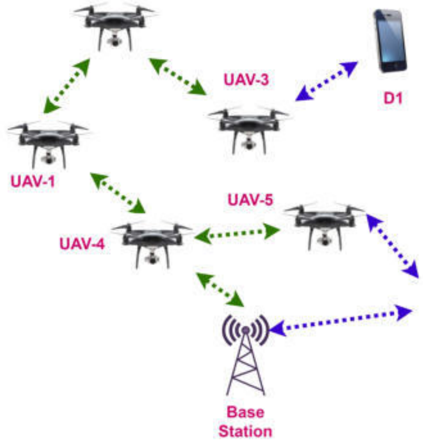


Fig. 1.1: FANET with IoT Communication

fuzzy logic to make decisions related to various aspects of the network [4]. To offer immediate communication in a challenging and distracting setting, FANET applications have been extended to other networks in the past several years. While FANETs have many potential uses, distributed optimization becomes more difficult due to the unreliability of connectivity caused by the increased mobility of UAVs [5]. Prior studies have focused on improving flying ad hoc communication and Internet of Things (IoT) equipped devices to address future difficulties before using them in the real world. A few of them fix problems with routing, resource utilization, predicting the movement of unmanned aerial vehicle (UAV) nodes, deciding on the topology of the network, security, and quality of service. Researchers gather all relevant issues and potential solutions, as well as write the research statement, before proceeding with the project. Fuzzy Methods for FANET and Internet of Things Congestion Avoidance in a methodical approach, this issue is resolved [6] [7] [8]. At first, it assesses the capacity of each node in the network in terms of all related metrics, such as energy consumption, processing speed, memory acquired for data buffering, channel bandwidth across links, and so on. At the moment of routing decision, when communication is being formed and initiated, all of the nodes are ranked from highest to lowest based on the aforementioned factors.

The article is divided into six sections. section 1 describes about introduction of FANET, and congestion control using fuzzy rule, section 2 gives a detailed explanation of existing work of congestion control in FANET, section 3 describes the proposed fuzzy logic technique for congestion control. Section 4 describes the proposed algorithm for congestion control, and section 5 shows the outcome of the proposed work and compares it with the existing system of congestion control in the last section 6 describes the conclusion and future work of the research article.

2. Issues in FANET and IOT. FANET and the Internet of Things (IoT) are two fast emerging technologies that, when coupled, offer a plethora of possibilities and applications. However, its integration poses many obstacles and issues [1][4][9]. Here are some of the major difficulties raised by integrating FANET and IoT.

Reliability of communication. FANET is based on wireless communication between aerial nodes or drones. Interference, signal attenuation, and limited bandwidth can all have an impact on communication reliability. A dependable communication link is critical in IoT applications, particularly those demanding real-time data delivery.

Scalability of a network. The network must scale as the number of IoT devices and aerial nodes grows. Managing a large-scale FANET with numerous networked IoT devices involves network architecture, addressing, and efficient routing problems.

Constraints on Energy. Drones in FANET and IoT devices frequently have low battery power. Energy-efficient communication protocols and solutions for regulating the energy consumption of aerial nodes and IoT devices are crucial for network operations to continue [10].

Privacy and security. Security is a major issue in both FANET and IoT. Combining the two creates new issues in terms of safeguarding communication channels, protecting data integrity, and ensuring the privacy of data collected by IoT devices.

Topology of a Dynamic Network. Because of the mobility of drones, FANETs have dynamic and unpredictable network topologies. Integrating IoT devices in such a dynamic environment necessitates adaptive routing algorithms and protocols capable of dealing with frequent topology changes [11].

Coordination and synchronization. Maintaining synchronization and coordination across drones and IoT devices is critical for data collecting, processing, and decision-making efficiency. Synchronization issues exist in a dynamic and decentralized FANET, and they must be solved.

Regulatory and Legal Considerations. FANET and IoT integration may generate regulatory and legal problems about airspace control, data ownership, and privacy. Following existing regulations and ensuring compliance with new ones becomes critical.

Fusion and processing of data. FANET generates massive amounts of data, which IoT devices contribute to. To extract relevant information from the merged data sources in real time, efficient methods for data fusion, processing, and analytics are required.

Autonomous mode of operation. FANET and IoT systems frequently operate autonomously or semi-autonomously. It is vital to ensure the reliability and safety of autonomous systems, especially in dynamic and unpredictable contexts.

Environmental Implications. The environmental impact of FANET, including drone energy consumption and electronic trash disposal from IoT devices, should be considered. The design and operation of these systems must incorporate sustainable practices and technologies.

To address these difficulties, professionals in communication systems, control theory, cybersecurity, and regulatory compliance must work together. Furthermore, continual research and development are required to establish solid solutions for FANET and IoT integration.

3. Literature Survey. This section aims to enhance understanding of current advancements in the areas of congestion, energy efficiency, and load balancing. The authors proposed numerous strategies for mitigating congestion in FANET with IoT. The recent work of authors considered.

Nousheen Akhtar et al. [12], proposed a bandwidth ware routing scheme (BARS) that is sensitive to bandwidth. It caches information in a queue to dynamically modify communication rates and alleviate congestion. The technique enables the source to modify its transmission rate whenever the network is close to experiencing congestion. We adapt the existing AODV protocol based on the available bandwidth in the network and the remaining queue sizes of each node in the path. The suggested routing strategy alters the RREQ and RREP messages of AODV by incorporating information about path bandwidth and queue size. Furthermore, the RERR message is also adapted to address path disconnection. To ensure high-quality routing, we have employed bandwidth and queue size as criteria for selecting routes. The limitation of this research is PDR decreases when mobility increases and packet loss % is more than 16%.

Manjit Kaur et al. [13], proposed a highly effective load balancing algorithm in FANET. This work is based on the traffic congestion control algorithm as a problem of optimizing network utility, while considering several network characteristics. The suggested method distributes the computing load among airborne nodes while determining the location of unfamiliar nodes. Furthermore, the method has been enhanced by integrating the Firefly algorithm and the traffic congestion control algorithm into a FANET. The limitations of the research include a simulation time of only 12 milliseconds and a recommended technique with a PDR value above 100, which is not feasible. The analysis of packet dropping resulting from congestion is absent.

Shaojie Wen et al. [14], proposed an Optimization of distributed systems with time constraints in Flying Ad-Hoc Networks (FANETs) using both primal and dual decompositions." The objective of this title is to enhance several network characteristics in a decentralized manner for delay-constrained flying ad hoc networks (FANETs) without having access to global network topology information. For this purpose, every Unmanned Aerial Vehicle (UAV) calculates the mean amount of disturbance over a specific duration to ascertain the status of the channels. The distributed optimization problem is thus expressed as a utility maximization problem that simultaneously optimizes power control, rate allocation, and routing with constraints on delay. A method is shown to eliminate the restriction on connection capacity, employing a dual approach. The primary limitation of the research is to analyze the varying speeds of UAVs, which are influenced by communication factors and the results lack an overhead analysis.

Lingli Yang et al. [15], proposed a RES-TDMA, which is a decentralized scheduling system based on time-division multiple access (TDMA), specifically designed for FANET. This protocol can allocate time slots dynamically by monitoring packets that request time slot reservations. By managing the traffic table, network nodes can obtain time slot allocation information within a maximum of two hops, as well as promptly recognize and free up time slots. Furthermore, the integration of an intent-driven network into FANET, the proposed RES-TDMA, provides the ability to perform self-analysis and self-configuration functions depending on the traffic intent. Our proposed technique effectively reduces the influence of node mobility and enhances the utilization of time slots. The primary limitation of the research is the absence of an investigation of the specific number of operational UAVs within a given time window. The impact of time slots on packet reception is not assessed and the time slot method for long-term users is not specified.

G. Soni et al. [16], proposed a novel privacy-preserving under dense traffic management (PPDM) routing method to safeguard the 6G-VANET from malicious black hole attacks in VANET. Black hole cars disregard essential information from traffic status packets transmitted by leading vehicles to trailing vehicles. A security system is capable of detecting and effectively preventing the loss of data packets within a network node. The existing SAODV security system is compared to the innovative PPDM system. By implementing a ban on aggressive cars within the network, the PPDM effectively safeguards and improves the performance of the VANET. An evaluation is conducted to compare the effectiveness of the proposed PPDM system with the existing SAODV. The PPDM demonstrates superior performance and less data loss as compared to the SAODV. The primary limitation of the research is its exclusive focus on detecting a single node, rather than several nodes. The drop in performance is solely attributed to the attacker not being assessed and the role of RSU in assault detection is not elucidated. Torkzadeh et al. [17], proposed a distinctive and efficient evolutionary method to tackle this problem. The researchers introduced an innovative QoS routing algorithm that incorporates evolutionary approaches (EAs). This algorithm is efficient and generates feasible solutions within a brief timeframe. The goal of the EAs is to ascertain the best appropriate and feasible solution for the given circumstance. In order to accomplish this, we initially assess the criteria of our problem, specifically the task of determining a viable route from a designated starting point to a specified endpoint inside an extensive network. The main goal of routing algorithms is to choose a path in a flexible, intelligent, and adjustable fashion. The primary limitation of the research is the minimal disparity in the success rate between the suggested strategy and earlier approaches and what is the benefit of this extremely low success rate? It is not specified. Assessment of routing performance with respect to data packets is lacking.

Sharma et al. [18] proposed a Distributed priority tree-based routing algorithm for FANETs. Their study focused on network partitioning between aerial and ground ad hoc networks and aimed to build a routing protocol that can effectively handle transmission in a coordinated system. The system relies on a combination of three primary criteria: link quality, traffic load, and spatial distance. Pu et al. [19], proposed a multipath routing protocol specifically designed for flying ad hoc networks (FANETs) to mitigate intentional jamming, disruption, isolated failures, and localized failures. The aim is to prevent these issues from negatively impacting the overall network performance of FANETs. Fang et al. [20] proposed a hybrid media access control mechanism for aeronautical ad hoc networks that ensures quality of service (QoS). The protocol is based on pre-allocating transmission time slots and providing rapid access. The aforementioned routing algorithms exclusively consider transmission dependability and disregard the constraint of packet latency.

M. Ploumidis et al. [21], proposed a method for allocating flows in random-access wireless multi-hop networks. The goal of their approach is to maximize throughput and limit latency by allocating different flows across numerous discontinuous pathways. This is particularly relevant for networks with multi-packet reception capabilities. In order to enhance the overall flow throughput and minimize packet latency, the issue is formulated as a non-convex optimization problem, and a distributed flow allocation method is suggested.

4. Proposed Approach. A flying ad hoc network is a collection of highly movable nodes to interconnect by wireless medium, which has a low capability of processing power, memory, and energy retention. Due to their limitations, it faces the congestion problem because it has a low bandwidth capacity. In the section of the existing survey, we study various congestion resolution algorithms that deal with overcoming the problem of congestion and improving the service quality of the network. In this article, our objective is to implement a fuzzy rule-based congestion control technique for flying ad hoc networks (MCPBF), which detects and further avoids congestion from the network as compared to the existing approach. The section describes how the fuzzy rules work, what parameters are taken to control and avoid congestion in the network, and the type of output impact after the avoidance rule is applied.

Assuming we want to determine the level of congestion in the FANET based on factors such as channel utilization, data interval, energy and memory utilization. Flying ad hoc networks form the route decision in a dynamic way, which takes time complexity O(n2), and after the route establishment process, the source device sends data to the base station or receiver node, which requires time complexity O(n) for data transmission. Analyzing Memory, Data interval, Channel, and Process Utilization for Fuzzy-Based Congestion Detection and Avoidance in Flying Ad Hoc and IoT Networks (MCPFB) entails assessing how the proposed system manages and uses these resources. Here's how may go about approaching this analysis:

4.1. Data Interval. In data transmission between nodes, the data interval often refers to the time period between data packets. The channel bandwidth, data size, and data type all affect the amount of time between data packets. In network communication, when the data interval is lower than the average interval, congestion occurs; on the other hand, when the interval is higher than the average data interval, bandwidth utilization is low, necessitating the use of a technique that keeps the data interval consistent. To preserve the consistency of data intervals, MCPBF first identifies the data interval and obtains a fuzzy inference such as low, medium, or high, and then applies the congestion avoidance approach when the data interval becomes low and leads to congestion.

4.2. Energy Utilization. Flying devices have a low capability of energy devices because they only work under limited battery power, which has the chance to increase the sudden communication loss. To improve communication, it's more important to utilize energy resources in an efficient way which is possible to increase the communication time using low energy utilization. Energy parameters are involved in calculating the congestion level when a node has low energy and sudden loss of network which changes the route from one to another and increases the load of other paths, so in the proposed MCPBF approach more emphasis on detecting every node energy to be aware of congestion level. Congestion is detected using a fuzzy-based technique which uses the linguistic variable as (Low, Medium, High) their rules if defined in below section.

4.3. Memory Utilization. Flying ad hoc devices having low capable memory units, which handle the incoming and outgoing data flows of devices. While the incoming data flow of any intermediate device is higher than the outgoing then it increases the utilization of memory of the device and any instance of time node memory is fully utilized which raises the problem of congestion. The memory of the device is indirectly dependent on congestion which is monitored by a fuzzy rule-based technique their linguistic variables are (Low, Medium, and High).

4.4. Channel Utilization. Congestion occurs in a flying ad hoc network when each node delivers data to its destination at the same time, exceeding the capacity of the channel. At the moment, just the data link layer is active, allowing the source nodes to perceive the medium all the way to the next linked node. However, when the capacity of that shared connection is depleted due to severe demand, congestion in the network becomes an issue. The rules of the suggested (MCPBF) method, which uses a fuzzy-based approach to identify and avoid network congestion, are defined in the section below. By systematically analyzing these aspects, you can gain insights into how the MCPFB system performs in terms of memory, channel, and process utilization for fuzzy-based congestion detection and avoidance in a Flying Ad Hoc and IoT Network.

Rule 1: IF (Data interval is Low) AND (Channel Uses is High) AND (Memory Uses is High) **THEN** (Congestion is High)
Rule 2: IF (Data interval is Medium |High) AND (Channel Uses is Medium) AND (Memory Uses is Medium) **THEN**

Table 5.1: Parameters for simulation

Parameters	Configuration Value
Simulation Tool	NS-2.31
Routing Protocol	BARS, MCPBF, WRA, DRA
Simulation Area	1650m∗1065m
Network Type	FANET
Number of Nodes	69
Physical Medium	Wireless, 802.11
Simulation Time (Sec)	300Sec
MAC Layer	802.11
Antenna Model	Omni Antenna
Traffic Type	CBR, FTP
Propagation radio model	Two ray ground
Energy (Initial)/J	Random

(Congestion is Medium)

Rule 3: IF (Data interval is Medium |High) AND (Channel Uses is Low) AND (Memory Uses is Low) **THEN** (Congestion is Low)

4.5. Proposed MCPFB Algorithm. This section describes the formal description of memory, channel, and process utilization to detect congestion using a fuzzy rule-based method which is classified into three levels i.e. low, medium, and high. We get high congestion that is resolved by minimizing channel utilization and memory utilization methods and increasing data interval to overcome the congestion in the flying ad hoc network. The algorithm will be divided into three parts: input, procedure, and output. In the input section, declare the variables such as source device, receiver device, fuzzy variables, data interval, protocol type, data type, etc. All these variables are configured with NS-2.31 by calling the function using the declared variable. In the procedure section, perform the routing procedure using data interval, channel utilization, and memory utilization of each intermediate flying device, and apply a fuzzy rule to select the best path, which is a congestion-free route; the other path is under the category of heavy or medium congestion status, so that the nodes are not selected for communication, and resolve the issue of congestion by data rate minimization, etc. In the output section, retrieve the results of the MCPFB algorithm in terms of throughput, packet delivery ratio, routing load, congestion status, energy, memory, channel utilization status, etc. with the help of the proposed algorithm to detect and avoid congestion in the IoT-flying ad hoc network.

5. Simulation Parameters. Simulation parameters depend on the specific context and goals of your simulation. However, for a general simulation involving Free-floating Aerial Networks (FANET) with Internet of Things (IoT) devices. The simulation parameters considered for simulation are simulation time that should be sufficient to capture relevant events and behaviors, the number of nodes, the communication range, and the rest of the parameters mentioned in Table 5.1.

6. Result Description. This section mentions the result analysis of previous BARS and the proposed MCPFB approach. The performance of both the protocols evaluated by performance metrics and the performance of MCPFB is better.

6.1. Throughput Analysis. Throughput is the amount of data sent from one end to the other in a certain amount of time. If the received packets are of higher quality, there will be a longer delay in data retransmission. This implies that the throughput parameter is good for measuring packet receiving at the destination end because there is no room for data retransmission and the network is congested, it is desirable to have a significant delay in successful transmission. In this graph, we have taken the analysis of DRA, WRA, BARS, and MCPFB protocols, where MCPFB has a maximum throughput performance of 550 kbps, BARS has a maximum throughput performance of 400 kbps, WRA has 400 kbps, and DRA's maximum throughput is nearly 340 kbps. With the comparative analysis of throughput, we conclude that the proposed MCPFB

Algorithm 1 MCPFB

Input:
N_t**: Network type IoT-FANET**
f_d**: flying device**
D_s**: Data source device** ϵ f_d
R_d**/BTS: Data receiver device or base receiver** ϵ f_d
I_f**: Intermediate flying device** ϵ f_d
$F_z v$**: (L, M, H) fuzzy value**
Ch_u**: Channel utilization**
M_{uti}**: Memory utilization**
D_{inter}**: Data Interval**
E_{uti}**: Energy utilization**
$Cong_s$**: Congestion Status**
R_{pt}**: routing protocol MCPBF**
D_{type}**: Data type TCP, UDP**
Ψ**: radio range** $550m^2$

Output:
Throughput, Packet Delivery Ratio, Routing Overhead, Congestion Status, Data interval, Energy, Memory utilization.

Procedure:

```
Form $N_t$ with active $f_d$
$D_s$ want to sent data to $R_d$
$D_s$ call $R_{pt}$ and Create packet ($D_s$, $R_d$, $R_{pt}$)
  while (visited $\neq f_d$ OR $I_f \neq R_d$) do
    if ($I_f$ in $\Psi$ and $I_f \neq R_d$) then
      Calculate ($D_{inter}$, $E_{uti}$, $Ch_u$, $M_u ti$) of $I_f$
      Apply MCPBF for fuzzy inference
      if (($D_{inter}$ is Low) AND ($Ch_u$ is High) AND ($M_u ti$ is High) then
        $I_f$ ($Cong_s$ ) $\leftarrow$ High
        $I_f$ $\leftarrow$ node not selected
        $I_f - 1$ forward route packet to other next-hop
        $I_f = I_f + 1$
      else if ($D_{inter}$ is medium) AND ($Ch_u$ is medium) AND ($M_{uti}$ is medium) then
        $I_f$ ($Cong_s$ ) $\leftarrow$ Medium
        $I_f$ $\leftarrow$ Selected in route and $I_f$ stop receiving new route packet
        $I_f$ $\leftarrow$ forward route packet to next-hop
      else
        $I_f$ ($Cong_s$ ) $\leftarrow$ low
        $I_f$ $\leftarrow$ Selected in route and $I_f$ stop receiving new route packet
        $I_f$ $\leftarrow$ forward route packet to next-hop
        $I_f = I_f + 1$
      end if
    else if ($I_f$ in $\Psi$) and ($I_f == R_d$ ) then
      $I_f \leftarrow l_f$
      $R_d$ receiver route packet
      if (Path > 1) then
        Calculate $E_{uti}$ of each node in $\forall$ paths‘

        if $path_i(E_{uti}) < path_j(E_{uti})$ then
          $Path_i$ select for communication
        else
          $Path_j$ select for communication
        end if
      end if
      $R_d$ Send acknowledgement to $D_s$
      $D_s$ call $Data_{pkt}$ ($D_s, R_d, D_{type}$)
    end if
    $Data_{pkt}$ ($D_s, R_d, D_{type}$)
    $D_s$ start data sending to $R_d$
    Check $Cong_s$ of each $I_f$ node in path
    if ($D_{inter}$ is Low) AND ($Ch_u$ is High) AND ($M_{uti}$ is high or medium) then
      $I_f$ ($Cong_s$ ) $\leftarrow$ High or Medium
      Increase $D_{inter}$ or min($P_{size}$)
    else
      $I_f$ ($Cong_s$ ) $\leftarrow$ Low
      $D_s$ send $D_{type}$ without changing $P_{size}$ and $D_{inter}$
    end if
  end while
```

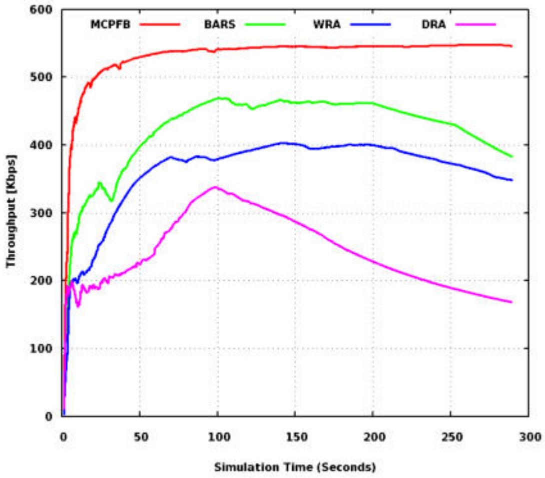


Fig. 6.1: Throughput Analysis

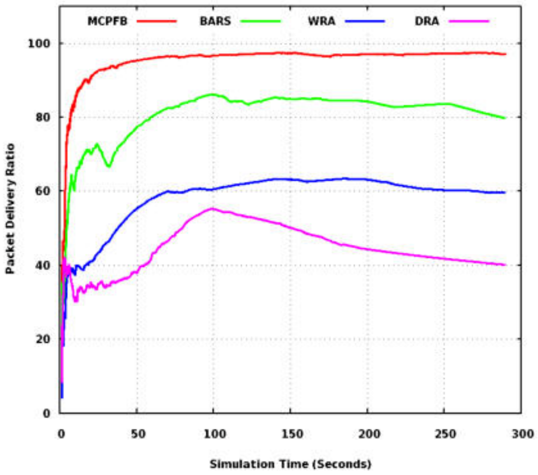


Fig. 6.2: PDR Analysis

technique gives efficient performance with a congestion-free network.

6.2. Packet Delivery Ratio Analysis. The data-receiving percentage performance indicates how much data was successfully received and delivered to the destination. The available bandwidth capacity is the most crucial feature of any wireless connection, and having adequate bandwidth means minimal data loss. Initially, the MCPFB technique in the network establishes paths through nodes that get a high signal strength. The MCPFB has received 92%, when the BARS have received 80% and two other existing techniques WRA and DRA packet delivery ratio less than 60%. At the destination end, MCPFB ensures that maximum number of packets is successfully received. The MCPFB technique allows for effective channel utilization as well as optimal bandwidth utilization and energy utilization. When network congestion is addressed appropriately using the provided technique, data loss is reduced and performance is improved.

6.3. Routing Overhead Analysis. The normal routing load is defined as the ratio of the number of data packets received to the number of packets utilized to establish connections. As the number of greeting or control packets increases, so does the amount of bandwidth consumed. This indicates that when the bandwidth

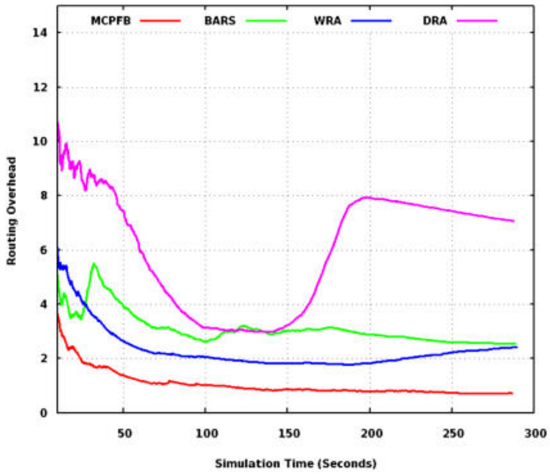


Fig. 6.3: Routing Overhead Analysis

Table 6.1: Data interval analysis

	BARS		MCPFB	
Connection Pair	Data Interval	Status	Data Interval	Status
0< − >60	0.16	Medium	0.32	Low
10< − >61	0.1	High	0.166666667	Medium
15< − >62	0.3	Low	0.333333333	Low
20< − >63	0.2	Medium	0.222222222	Low
35< − >3	High			
45< − >65	0.12	High	0.171428571	Medium
47< − >66	0.21	Low	0.35	Low
50< − >67	0.51	Low	0.566666667	Low
52< − >68	0.19	Medium	0.211111111	Low

utilized by hello packets is less than the bandwidth available for data packets. The NRL of previous technique WRA, DRA and BARS is higher than MCPFB approach in FANET-IoT network. The NRL of proposed MCPFB is less than 1, it is barely 2 in the beginning of simulation. The overhead of BARS is three times more which shows unnecessary bandwidth consumption. When there is less overhead, the link between the sender and the recipient becomes more reliable. The amount of data lost in the network is lowered as a result of the decreased possibility of packet retransmission. The only way to accomplish this extraordinary result is to use buffer management and multipath path routing in tandem with suitable channel allocation.

6.4. Data Interval Analysis. The data interval analysis uses the fuzzy linguistic (low, medium, and high) technique mentioned in Table 6.1, and the status of MCPFB congestion is low on some connection pairs as compared to the previous BARS scheme in the network. The connection pair 0 and 60, 45 and 65 or more congestion status has been approved. Less congestion means more fast packets receiving.

6.5. Channel Analysis. The channel analysis using the fuzzy linguistic (low, medium, and high) technique is mentioned in table 6.2 and the status of MCPFB channel utilization is showing medium status but that was high in the previous. The connection pairs 0 and 60, 20 and 63 utilize channel efficiently. Proper channel utilization gives better as compared to MCPFB.

6.6. Queue Analysis . The queue analysis is mentioned in Table 6.3 and the status of MCPFB congestion is medium to low for connection 45 to 65. The queue utilization shows balanced data forwarding in the network.

Table 6.2: Channel Analysis

	BARS		MCPFB	
Connection Pair	Channel Utilization	Status	Channel Utilization	Status
0< − >60	100	High	50	Medium
10< − >61	100	High	66	High
15< − >62	40	Medium	35	Medium
20< − >63	75	High	35	Medium
35< − >64	100	High	100	High
45< − >65	42	Medium	34	Medium
47< − >66	42	Medium	40	Medium
50< − >67	3	Low	3	Low
52< − >68	15	Low	13	Low

Table 6.3: Queue Analysis

	BARS		MCPFB	
Connection Pair	Queue Uses	Status	Queue Uses	Status
0< − >60	16	Medium	14	Medium
10< − >61	20	Medium	18	Medium
15< − >62	24	Medium	22	Medium
20< − >63	3	Low	3	Low
35< − >64	4	Low	4	Low
45< − >65	10	Medium	9	Low
47< − >66	18	Medium	16	Medium
50< − >67	4	Low	4	Low
52< − >68	6	Low	5	Low

Table 6.4: Energy Utilization Analysis

	BARS		MCPFB	
Node Pair	Percentage of Utilization(E)	Status	Percentage of Utilization(E)	Status
0< − >60	38.46	Medium	30.59	Medium
10< − >61	50	High	45.44	High
15< − >62	42.1	High	33.93	Medium
20< − >63	42.34	High	32.13	Medium
35< − >64	39.72	Medium	28.69	Medium
45< − >65	50	High	45.22	High
47< − >66	47.11	High	33.15	Medium
50< − >67	37.04	Medium	29.81	Medium
52< − >68	35.1	Medium	25.36	Medium

The normal queue utilization means faster packet receiving.

6.7. Energy Utilization Analysis. The energy utilization of low-capacity devices is impacted by more parameters because if device energy utilization is higher, it means devices frequently change the route from a dead path to an alive path. In this section, describe in Table 6.4 the simulated analysis of energy utilization in the existing BARS and the proposed MCPBF technique using fuzzy linguistic variables (low, medium, and high) and conclude that the proposed MCPFB is efficient because it minimizes energy utilization as compared to the BARS technique. It means MCPFB minimizes energy consumption while increasing network stability.

Table 6.5: Summarized Performance Analysis

Parameters	BARS	MCPFB	WRA	DRA
Number of Packets Sends	14529	15866	15026	12966
Number of Packets Receives	11584	15388	8977	5208
Percentage of Data Receives	79.73	96.99	59.74	40.16
Normal Routing Load	5.07	1.43	2.41	7.06
Average e-e delay(ms)	149.29	66.53	73.36	190.12
Average Energy Consume	83.2	63.26	80.25	90.85
Average Residual Energy	16.46	36.1	19.28	8.88

6.8. Summarized Performance Analysis. In this section, we describe the summarized results of FANET-IoT application protocols. Table 6.5 compares the performance of the existing technique DRA, WRA, and BARS with the proposed MCPFB and gets the outcome in terms of the number of packets sent, received, percentage of data received, routing overhead, delay, and energy consumption. All collective parameters perform more efficiently and are more adoptable while applying the proposed MCPFB technique, which is useful for futuristic FANET-IoT.

7. Conclusion and Future Work. In the Flying Ad-hoc Network (FANET), UAVs build a temporary network with IoT devices to send information from one location to another or to other devices, such as IoT. The UAVS has the ability to collect and record current state data and transfer it to distant IoT devices. The data interval, bandwidth, and memory space all play key roles in the proper communication of hybrid devices in networks. In this study, IoT devices were able to send direct instructions to any other IoT device via regular connectivity. In order to identify and avoid network congestion more effectively than the current approach, this study proposes a fuzzy rule-based congestion control technique known as MCPFB. The MCPFB explains how fuzzy rules work, the parameters used to manage and prevent network congestion, and the consequent impact on output after applying the avoidance rule. To determine the amount of congestion in the FANET, we will look at channel use, data interval, energy utilization, and memory utilization. Finding out how well the proposed system manages and makes use of these resources is part of studying Memory, Data Interval, Channel, and Process Utilization for Fuzzy-Based Congestion Detection and Avoidance in Flying Ad Hoc and IoT Networks (MCPFB). The UAVs cannot examine the data to decide if it is valid or erroneous. The bandwidth and processing capabilities of UAVs are always difficult for researchers to manage and send across FANET. The congestion control approaches mentioned in this study are trustworthy and capable of improving network performance. MCPFB aims to enhance accuracy and produce better outcomes than earlier approaches. The MCPFB throughput performance is more than 100 kbps greater than BARS, WRA, and DRA, and the PDR is more than 15%, indicating higher data packet reception. The better data receiving means lower overhead, which is why the overhead of MCPFB is 1.4, resulting in a less congested network than other existing techniques, and BARS is 5.1 (four times) higher than the other two existing WRA and DRA, which also have higher overhead, increasing network congestion. In the future, try to propose a fuzzy-based security approach in FANET against flooding. The role of the fuzzy rule is to detect the attacker's presence and the prevention scheme's role is to disable the exitance of an attacker to control the malicious actions in the network.

REFERENCES

[1] Jatin Sharma, Pawan Singh Mehra , *Secure communication in IOT-based UAV networks: A systematic survey*, Internet of Things, 2023, 23.

[2] Kamlesh Chandravanshi, Gaurav Soni, Durgesh Kumar Mishra , *Design and Analysis of an Energy-Efficient Load Balancing and Bandwidth Aware Adaptive Multipath N-Channel Routing Approach in MANET.*, IEEE Access, 2022, 10, pp. 110003 – 110025.

[3] T.S. Pradeep Kumar, M. Alamelu , *Modelling and Simulation of Fast-Moving Ad-Hoc Networks (FANETs and VANETs)* , IGI Global, 2022.

[4] Josh Howarth, *Amazon IoT Statistics*, 2023.

[5] Tanweer Alam, *Fuzzy Control Based Mobility Framework for Evaluating Mobility Models in MANET of Smart Devices*, ARPN Journal of Engineering and Applied Sciences, 2017, 12(15).

[6] R. Uma Mageswari R, Nallarasu Krishnan, Mohammed Sirajudeen Yoosuf, K. Murugan and C. Sankar Ram, *Establishment of FANETs Using IoT-Based UAV and Its Issues Related to Mobility and Authentication*, Modelling and Simulation of Fast-Moving Ad-Hoc Networks (FANETs and VANETs), 2023, Chapter-4, pp.74–93.

[7] G. Soni, K. Chandravanshi, A.S Kaurav, S.R Dutta, *A Bandwidth-Efficient and Quick Response Traffic Congestion Control QoS Approach for VANET in 6G*, Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems, 2022, 385, pp. 1–9.

[8] Armir Bujari, Carlos T Calafate, Juan-Carlos Cano, Pietro Manzoni, Flying Ad-hoc Network Application Scenarios and Mobility Models, International Journal of Distributed Sensor Networks, 2017, 13(10).

[9] G. Soni, M. K. Jhariya, K. Chandravanshi and D. Tomar, *A Multipath Location based Hybrid DMR Protocol in MANET*, IEEE 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), 2020, pp. 191–196.

[10] Gaurav Soni, R. Sudhakar, Krishna Pathak, *Cluster based Techniques for Eradicating Congestion in WSN Aided IoT: A Survey*, International Journal of Emerging Technology and Advanced Engineering, 2018, (8)12, pp. 166–172.

[11] Megala, V Kathiresan, *Novel Clustering Based Energy Efficient with Adaptive PSO Approach For Congestion Control in FANET*, Webology, 2021, 18(6).

[12] Nousheen Akhtar, Muazzam A. Khan, Ata Ullah, and Muhammad Younus Javed, *Congestion Avoidance for Smart Devices by Caching Information in MANETS and IoT*, IEEE Access, 2019, 7, pp. 71459-–71471.

[13] Manjit Kaur, D. Prashar, M. Rashid, Z. Khanam, S.S Alshamrani, A. S Al Ghamdi, *An Optimized Load Balancing Using Firefly Algorithm in Flying Ad-Hoc Network*, Electronics, 2022.

[14] Shaojie Wen, Lianbing Deng, Yuhang Liu, *Distributed optimization via primal and dual decompositions for delay-constrained FANETs*, Elsevier Ad Hoc Networks, 2020, 109.

[15] Lingli Yang, Peilin Tao, Tong Li, Chungang Yang, Xinru Mi, Ying Ouyang, Donghong Wei, Qicai Wang, *Reservation and Traffic Intent-Aware Dynamic Resource Allocation for FANET*, IEEE 21st International Conference on Communication Technology (ICCT), Tianjin, China, 2021, pp. 971-975.

[16] G. Soni, K. Chandravanshi, *A Novel Privacy-Preserving and Denser Traffic Management System in 6G-VANET Routing Against Black Hole Attack*, Sustainable Communication Networks and Application, Lecture Notes on Data Engineering and Communications Technologies, 2022, 93.

[17] S. Torkzadeh, H. Soltanizadeh, A.A. Orouji, *Multi-constraint QoS Routing using a Customized Lightweight Evolutionary Strategy*, Soft Computing, 2019,23, pp. 693-–706.

[18] V. Sharma, R. Kumar, N. Kumar, *DPTR: Distributed priority tree-based routing protocol for FANETs* , Computer Communication, 2018, 122, pp. 129-–151.

[19] C. Pu, *Jamming-resilient multipath routing protocol for flying ad hoc networks*, IEEE Access, 2018, 6, pp. 68472-–68486.

[20] Z. Fang, Q.M. Qiu, Y.F. Ding, L.H. Ding, *A QoS guarantee based hybrid media access control protocol of aeronautical Ad hoc network* , Wirel. Personal Communication, 2018, 6, pp. 5954-–5961.

[21] M. Ploumidis, N. Pappas, A. Traganitis, *Flow allocation for maximum throughput and bounded delay on Multiple Disjoint Paths for Random Access Wireless Multihop Networks*, IEEE Transaction of Vehicular Technology, 2017,66(1), pp.720–733.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Mar 14, 2024
Accepted: May 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 361–370, DOI 10.12694/scpe.v26i1.3533

GENOME SEQUENCE ANALYSIS OF SEVERE ACUTE RESPIRATORY SYNDROME USING GENOANALYTICA MODEL

SHIVENDRA DUBEY[*], DINESH KUMAR VERMA[†], AND MAHESH KUMAR[‡]

Abstract. We proposed a GenoAnalytica model for examining the SARS's genomics sequences. The technologies make proper data extraction from genomics sequences of viruses. We use the GenoAnalytica model, i.e. GenoCompute, and IGMiner Algorithm; to classify the range of genomics sequences, including recognizing the sequence variation from the datasets. The projected algorithm computes the nucleotide patterns and represents the nucleotide genome sequence of SARS (airborne virus) by IGMiner technique and works out on the GenoCompute to calculate computation time with minimum count in second. Along with this, we proposed a UMRA algorithm to compute the mutation rate of the genome sequence with minimum count in seconds as compared to traditional method. Furthermore, we work out the different datasets (China and Algeria datasets) and determine the whole variation at the index level inside the all genome sequence. This learning also signifies the performance evaluation on altering minsup using IGMiner and Aprori-based SPM. Also, we calculate the mutation rate of the genome sequence of airborne virus using Unique Mutation Rate Analysis algorithm. The severe acute respiratory syndrome coronavirus 2 has been responsible for the deadly COVID-19 pandemic. It has ruined limitless individuals all over the globe, and along with this, it continues to harm well-being and people's health. Healthcare specialists and Researchers can obtain insight into COVID-19's inherited variation or SAR-CoV-2 through cutting-edge Artificial Intelligence and genome sequence analysis tools.

Key words: IGMiner, UMRA, Genome Sequence, COVID-19, GenoCompute

1. Introduction. Corona, also called COVID-19, represents a severe respiratory disease brought about by a new corona virus named SARS-CoV-2. The infectious disease was initially discovered in December 2019 in the Chinese city of Wuhan; it since then has spread around the world, causing a global epidemic. When a person with the infection of sneezes, coughs, breathes or speaks loudly, COVID-19 typically spreads by droplets from their lungs [1]. The virus may also be transferred by contacting infected surfaces or items, especially the mouth, face, eyes, or nose. The moderate to severe COVID-19 symptoms include coughing, fever, exhaustion, shortness of breath, muscular or body pains, loss of smell or taste, sore throats, and digestive and headaches problems. In extreme circumstances, it can result in mortality, organ failure, ARDS (acute respiratory distress syndrome), and pneumonia, particularly among older persons and people with underlying medical issues [2, 3]. It is crucial to practise excellent hygiene to stop the spreading of COVID-19. Examples include: washing your hands frequently with water and soap for at least twenty seconds; using a hand sanitizer with a minimum alcohol content of 60%; wearing masks in public; engaging in social distancing; avoiding big gatherings. In several nations, vaccines have been created and are authorized for use during situations of crisis to help prevent COVID-19. To guarantee these vaccinations' effectiveness and safety, they undergo extensive testing. The SARS-CoV-2 genome is an RNA genome, meaning RNA (ribonucleic acid) is used to carry biological data [4, 5]. The virus's RNA with one strand genome generates several proteins required for interaction and replication with its host cells. Scientists have pinpointed the precise genetic makeup of SARS-CoV-2 thanks to the sequencing of the virus' genome. Researchers can follow the development and transmission of the virus's genome and better understand how it is evolving and adapting as time passes by comparing several virus genomes from various historical periods and places. Additionally essential to the creation of COVID-19 tests for diagnosis is genome sequencing. Scientists may create tests to identify the existence of the pathogenic virus in patient specimens

[*] Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna, Madhya Pradesh, India, 473226 (shivendrashivay@gmail.com).

[†] Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna, Madhya Pradesh, India, 473226 (dinesh.hpp@gmail.com).

[‡] Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna, Madhya Pradesh, India, 473226 (mahesh.chahar@gmail.com).

by focusing on some regions of the genome that contain the virus, especially its genes that generate the protein known as the spike, which is essential for viral entrance into host cells [6].

There some key contributions of this research as:

1. This article proposes the GenoAnalytica method that offers an innovative technique for sequence rule mining through the use of the effective data structure and the classes of equivalence.
2. This article performs in-depth tests on actual datasets to verify the efficiency enhancements of GenoAnalytica, illustrating its usefulness for massive sequence rule mining jobs.
3. This study recognizes a modest rise in memory usage as a cost of improved speed, offering important details on the approach's resource needs.
4. The enhanced efficiency of ERMiner is anticipated to be advantageous to a broad variety of uses which depend on successive rule mining, including bio informatics, website click stream analysis, and market basket analysis.

2. Related Work. A novel optical biosensor [10] developed and integrates thermal and optical properties to identify the coronavirus. Mostly gold nanoislands connected on an optical substrate make up the sensor. Coronavirus RNA sequences corresponding to synthetic DNA receptors have been discovered on nanoislands. The sensor's receptors work in conjunction with the visible virus. Localized Surface Plasmon Resonance is the name of this technique (Swiss Federal Laboratories for Materials Science and Technique). A brand-new biosensor built around specific cells with altered mammalian membranes was proposed [11]. By including 10% foetal bovine serum, the author examined the circumstances of Green Monkey's renal culture of cells.

Additionally, Vero/Anti-S1 membrane-engineered cells were used to fabricate sensors. This cutting-edge biosensor proved that it could be used for COVID-19 antigen surface large-scale screening in about 3 minutes and produced a remarkable outcome. Utilizing two approaches, including the ANFIS (adaptive network-based fuzzy inference system) and multi-layered perception modelling, suggested a comparative study between the soft computing and machine learning models for predicting a global epidemic disease, highlighting the possibility of using machine learning as a tool for solutions in healthcare [8]. Technological devices offer enormous potential for interacting with mental health. Wristbands, Cellphones, and Smartwatches are examples of wearable devices having integrated sensors that can communicate through Wi-Fi or Bluetooth. Gyroscopes and accelerometers are examples of sensors that sense inertia. The sense of human body heavily relies on biological sensors that measure heart rate and environmental sensors that measure temperature. To address the issues of identity, trust, and privacy IJPCC developed [12] a ubiquitous system for computing that utilizes the trust model. The framework used naive Bayes (NB) and Apriori models to extract behaviour patterns during the decision-making process. For a summary of the advancements and research on wearing biomarker structures, particularly for monitoring one's health, wearable biosensors demonstrated [13]. Using smartphones presented a method to identify human behaviour that includes health-related behaviours, including physical activity for sleeping and fitness activity. The recommended concept uses sensors contained within the Arduino GNO as input to power systems that run computers. The COVID-19 illness is forecast-ed and categorized for subsequent therapy using artificial intelligence technology and a model based on mathematics [14].

3. Material and Methods.

3.1. Nucleotide Sequences. The exact sequence of the nucleotide in a molecule of RNA or DNA is referred to as a nucleotide sequence. These molecules' fundamental components, known as nucleotides, are made from three primary parts: a phosphate group, a nitrogenous base, ribose in RNA and deoxyribose in DNA(a sugar molecule). Adenine (A), cytosine (C), guanine (G), and thymine (T) are each of the four distinct bases composed of nitrogen that may be discovered in DNA. Uracil (U) takes the role of thymine in RNA. The nucleotide sequence comprises the arrangement and order of these bases across the sugar-phosphate backbone [7].

The steps required to create proteins and other valuable compounds are included in the arrangement of nucleotide sequences, which also house the biological information of a living thing. It gives biological data to identify an individual's traits and characteristics [8].

3.2. Genome sequencing. Discovering the whole sequence of DNA of the genome of a living thing is a procedure called genome sequencing. It entails defining the arrangement of the bases that make up nucleotide (adenine, cytosine, guanine, and thymine) within the DNA molecular structure, which transmits a living thing's genetic information. Genome sequencing may be done using various technologies and methodologies, with newer, more efficient ones made possible by technological breakthroughs [9]. The WGS (whole-genome sequencing) and TS (targeted sequencing) were the two main methods for sequencing the human genome.

3.3. Datasets. We have implemented this methodology on hp laptop with i5 processor; and the genome sequence dataset of SARS-CoV–2 of USA and Algeria collected from the NCBI repository. The genome sequence files has been use in common file extensions, like FASTA or GenBank, to allow for simple analysis and exchange across investigators. An investigator has examined patterns over a period of time by using the information set's potential inclusion of information on the time and dates each collecting the samples. The data sets have details about the site and geographic coordinates where the specimen took place. Using this, one may monitor the virus's international spread. It is often possible to identify and get particular sequences from a dataset by using the accession number, which is a code that is usually given to that particular sequence. The sequence information is usually structured within a certain pattern and is frequently expressed employing the common nucleotide coding (A, T, C, and G).

3.4. Proposed Method. The genome sequence's computing time is computed using Algorithm 1's and provided information. We use m = 3 and k = 9 to calculate the minimizes to comprehend the GenoCompute Algorithm better. These values were chosen based on experience with the conventional validation set technique. The frequency distribution vector-based representations are produced using the same methodology as the minimizes for specific nucleotide sequencing. We refer to this technique as Minimize Vector for convenience. A sequence database (SQDB) and the minconf and minsup thresholds are inputs to IGMiner. It first does a single scan of the database to create all equivalence classes of rules of Size 1*1, or including a single item in a single entity and antecedent in the consequent. The LMS (left search) function is then invoked to execute left merges across all left equivalence categories to find more extensive rules.

Similarly, the RMS (right search) function performs right merges for appropriate equivalence classes. Although left merges are permitted after right merges, it should be noted that the RMS (right search) technique may produce some additional left-equivalence classes. Those equivalence classes are kept in the left store structure. The processing of these classes of equivalence is done in a separate loop. The IGMiner Algorithm then returns the collection of discovered rules.

Algorithm 1 GenoCompute Algorithm

```
Input:
minconf: Minimum Confidence Threshold;
SQDB: A sequence database
Output:
Set of Valid Sequential Rules
1. Let Store = Empty set
2. Let rules = Empty set
3. Scan SQDB once to calculate EQ, the set of all uniformity classes of rules of size 1*1
4. For each left uniformity class C1 in EQ, do
5. LMS(C1, rules)
6. End
7. For each right uniformity class C2 in EQ, do
8. RMS(C2, rules, Store)
9. End
10. For each left uniformity class C3 in Store, do
11. RMS(C3)
12. End
13. Return rules
```

Each site in the COVID-19 sequences is compared to its appropriate place in the standard genome as part of the UMRA method, which determines the rate at which mutations occur. The total variety of modifications

Algorithm 2 IGMiner Algorithm

```
Input: Sequence s and integer k and m
Output: Set of Minimizes
minimizes = 0
def find_minimizes(sequence, k_length, m_length):
minimizes = set()
queue = []
current_min_index = 0
for i in range(len(sequence) - k_length + 1):
kmer = sequence[i:i+k_length]
queue.append(min(kmer[j:j+m_length] for j in range(k_length - m_length + 1)))
if len(queue) > m_length:
queue.pop(0)
if queue[-1] < queue[current_min_index]:
current_min_index = len(queue) - 1
if i >= k_length - m_length:
minimizes.add(queue[current_min_index])
return minimizes
```

and the overall frequency of aligned locations are counted. After that, the mutation rate is determined by multiplying the alteration score by the number of aligned positions. This approach can be enhanced and modified depending on the needs and demands for further study.

Algorithm 3 UMRA Algorithm (Unique Mutation Rate Analysis)

```
def calculate_mutation_rate(reference_sequence, covid_sequences):
mutation_count = 0
aligned_positions = 0
for i in range(len(reference_sequence)):
reference_nucleotide = reference_sequence[i]
for sequence in covid_sequences:
if i < len(sequence):
aligned_positions += 1
if sequence[i] != reference_nucleotide:
mutation_count += 1
mutation_rate = mutation_count / aligned_positions if aligned_positions > 0 else 0
return mutation_rate
# Example usage
reference_genome = ”ATCGATCGATCG...”
covid_samples = [”ATCGATCGATCG...”, ”ATCGGTCGATCG...”, ”ATCGATCGCTCG...”]
rate = calculate_mutation_rate(reference_genome, covid_samples)
print(”Mutation Rate:”, rate)
```

4. Results. The total length of the genomics order, the computing power available, and the level of difficulty of the statistical methods are some variables that can affect how long it takes to analyze the COVID-19 genome sequence. The explanation of the variables that might impact calculation time is as follows:

Genome Sequence Length. SARS-CoV-2, the virus that causes COVID-19, has a genome sequence of around 30,000 base pair combinations long. The sequence length will affect how long it takes to process and analyze the data. Sequence comparison and alignment algorithms could take longer as the Size rises.

Analysis Algorithms. The decision regarding the analysis algorithms significantly impacts computation time. There are many levels of computational complexity for various analytical methods, including variant

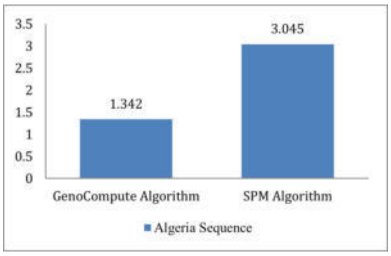


Fig. 4.1: Computation time of genetic sequences

calling, sequence alignment, and mutation detection. Several algorithms' implementations have been optimized to handle enormous datasets more quickly, cutting down on calculation time.

Computational Resources. Computation time may be significantly influenced by the available computational resources, such as the computer's processing speed and algorithm design. Analysis activities can be completed more quickly by dividing the workload across several processors or nodes into outstanding-performance cloud-based platforms or computing systems with concurrent processing capabilities. Here, we analyse the computation time of genetic sequences by two different algorithms which is shown in Figure 4.1.

Optimizing and paralleling. The effectiveness of the analytical methods and how they are implemented can impact computing time. Processing time can be decreased by optimized algorithms that minimize pointless operations or make use of effective data structures. The execution of analytical tasks concurrently can be made possible via parallelization techniques like distributed computing and multi-threading, which further reduces processing time.

Complexity and Size of the dataset. It might affect how quickly calculations are performed. The computational burden and processing time can be increased by analyzing excessive COVID-19 genomics sequences or incorporating more metadata, like clinical data or sample information. We worked on the Algeria sequences with GenoCompute algorithm and SPM algorithm (based method) to calculate the computation time, we get our proposed GenoCompute algorithm is much faster than SPM algorithm (see figure 4.1).

These variables significantly affect the calculation time, and it is challenging to give a specific time without considering the hardware resources, particular analytic pipelines, and dataset peculiarities. However, developments in parallel computing, computational biology, and algorithmic advances continue to increase computation time, the capacity for quicker analysis, and the effectiveness of genome sequence analysis of COVID-19 genome-wide information.

Knowing the COVID-19 dataset's genome sequence variation percentage offers essential information on the genetic variations of the virus's genome within a community or among many samples as shown in figure 2. Several factors make determining this proportion crucial, including the following:

Understanding Genetic Variability. The SARS-CoV-2 virus, the source of COVID-19, can demonstrate genetic changes or variations as it multiplies and spreads. The degree of biological variety within a dataset may be determined by computing the proportion of genome sequence diversity. This knowledge is essential for following the virus's development, spotting new variations, and comprehending their possible effects on pathogenic, transmission, or responsiveness to medications or vaccinations.

Monitoring the Spread and Transmission of Viruses. Analysis of Genome Sequence Variation (AGSV) enabled the tracking of viral transmission routes and the identification of virus clusters or lineages. We can evaluate the similarity of virus strains and monitor the spread of certain variations by comparing genomes from various places or periods. This data can help with contact tracing, public health actions, and tracking the success of control measures.

Genome sequence variation analysis supports epidemiological research and outbreak investigation. It assists researchers in comprehending the virus's genesis and dynamics of transmission, locating super-spreader events or clusters, and evaluating the effectiveness of treatments or containment measures. Sequences from various geographic or historical eras might be compared to recreate the virus' evolutionary history and learn more

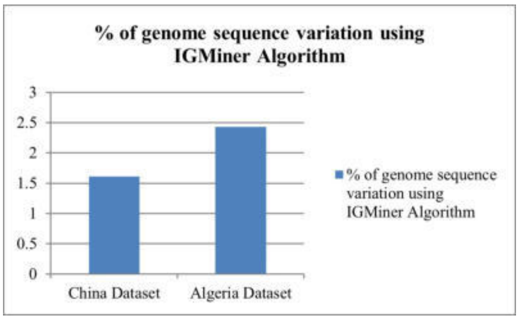


Fig. 4.2: Variation of genome sequence

Table 4.1: Variation in genome

Line	Location	Position	Change
1	23	23	T -> C
33	46	1966	T -> C
90	52	5392	G -> A
133	51	7971	A -> C
244	23	14603	T -> C
247	47	30305	A -> C

about how widely it has travelled.

We can investigate genetic variety, follow the transmission of the virus, evaluate the effectiveness of vaccines, and guide public health measures by measuring the proportion of the genome's sequence variations within COVID-19 databases. It is essential to expand our knowledge of this virus and enable decisions based on evidence in emergency response initiatives. Base method is not calculating percentage of genome sequence variation but our proposed IGMiner algorithm calculates this with high effectiveness (see figure 4.2).

There are several uses for determining the indexing of missing sequences in a COVID-19 dataset. Here are a few cases where deciding how to index missing sequences is crucial:

Evaluation of Completeness. A dataset's completeness can be inferred from the existence or absence of certain sequences. By determining missing sequences, researchers can evaluate whether particular geographic areas or viral strains are underrepresented or absent. Understanding the dataset's limits and any biases in the research is vital.

Comparative Analysis. In COVID-19 research, comparing the genetic sequences of several samples or areas is customary. The quality and thoroughness of comparison studies may suffer if specific sequences are absent from the dataset. Researchers can find deficiencies in the dataset and confidently decide whether the data are reliable and appropriate for particular analysis by computing the indices of missing sequences.

Identification of variations. Genetic variations or mutations can be identified by the absence or presence of specific sequences. Researchers can locate locations with novel or distinctive genetic variants by detecting missing sequences. Table 4.1 demonstrates the knowledge which can help with the detection and characterization of novel viral strains or variations in the population.

The integrity of Data and Control of Quality. One quality control tool is to determine how to index missing sequences. By eliminating any mistakes in data input or collecting, it helps to guarantee that the set of data is correct and complete. Researchers can preserve data integrity and improve the dataset's quality by checking for predicted sequences and locating missing ones.

Figure 4.3 Determining the order of indexing of missing sequences within the COVID-19 dataset enables researchers to judge the completeness of the dataset, ease comparison studies, pinpoint genetic variations, and guarantee quality control and data integrity. It allows for a more thorough and trustworthy comprehension of

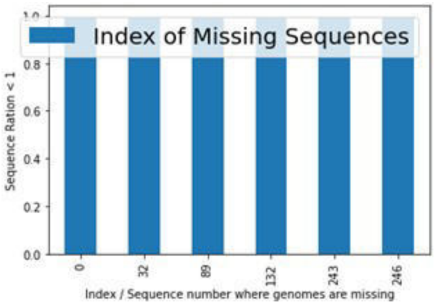


Fig. 4.3: Missing sequences

Table 4.2: Comparisons of Mutation with SPM-Point and GenoAnalytica Model

Dataset	SPM-Point Mutation (sec)	UMRA Model (sec)
China Dataset	3.96	2.73
Algeria Dataset	4.031	2.84

the virus and encourages reasoned judgement in public health and research activities.

We advise looking for information regarding the mutational study of the COVID-19 datasets in Algeria in reputable sources such as academic research papers, books from health organizations, or official materials provided by the Algerian Ministry of Healthcare or other authorities. In addition to any found variations or modifications in the COVID-19 virus in Algeria, these sites would offer current and reliable information regarding the mutation analysis, as shown in Figure 4.4.

It's crucial to understand that mutation analysis entails looking at the genetic structure of a virus and locating any alterations or variations in its molecular structure. This study aids in comprehending the genetic variety of the virus, monitoring the appearance and dissemination of various variations represented in Table 4.2, and evaluating the possible consequences for disease severity, transmission, therapies, diagnostics, and vaccinations. Here our proposed method worked on three data sets (USA, China and Algeria) to compute the mutations and we get our proposed method is so faster as compared to base method.

Figure 4.5 shows the comparison of support on various sequences with base method and IGMiner method. The time needed to identify linkages or correlations between objects in a dataset is called the ”association time.” The precise comparison relies on the data set's dimensions, complexity, chosen algorithm, and processing power. Direct comparisons of association periods are challenging without accurate information. However, the time required for association mining might vary greatly depending on the dataset's properties and methods. The particular dataset, including the amino acid sequences and their accompanying support values would typically be required to compare the support for sequencing found in the Algeria COVID-19 dataset. A simple comparison of support values for the genes in the Algeria COVID-19 dataset cannot be made with any accessibility to the actual dataset. We can, however, describe support in general and its importance in sequence analysis. Support in sequence analysis is the frequency or recurrence of an individual sequence pattern throughout a dataset. It reveals the relative frequency with which a specific design or sequence shows up in the dataset concerning the number of sequences. Support is essential for several data mining activities, such as sequential pattern mining and association rule mining. It assists in identifying frequent or significant sequences that are more likely to be noteworthy or instructive.

5. Conclusion. In this study, particular approaches for examining and investigating SARS-CoV-2 genomics sequences have been presented. The most important method, the pattern mining algorithm, recognizes frequent nucleotide bases inside the sequences and furthers their sequential and regular patterns. Numerous sequence prediction algorithms were experienced on genomics sequences. Furthermore, the outcomes point out

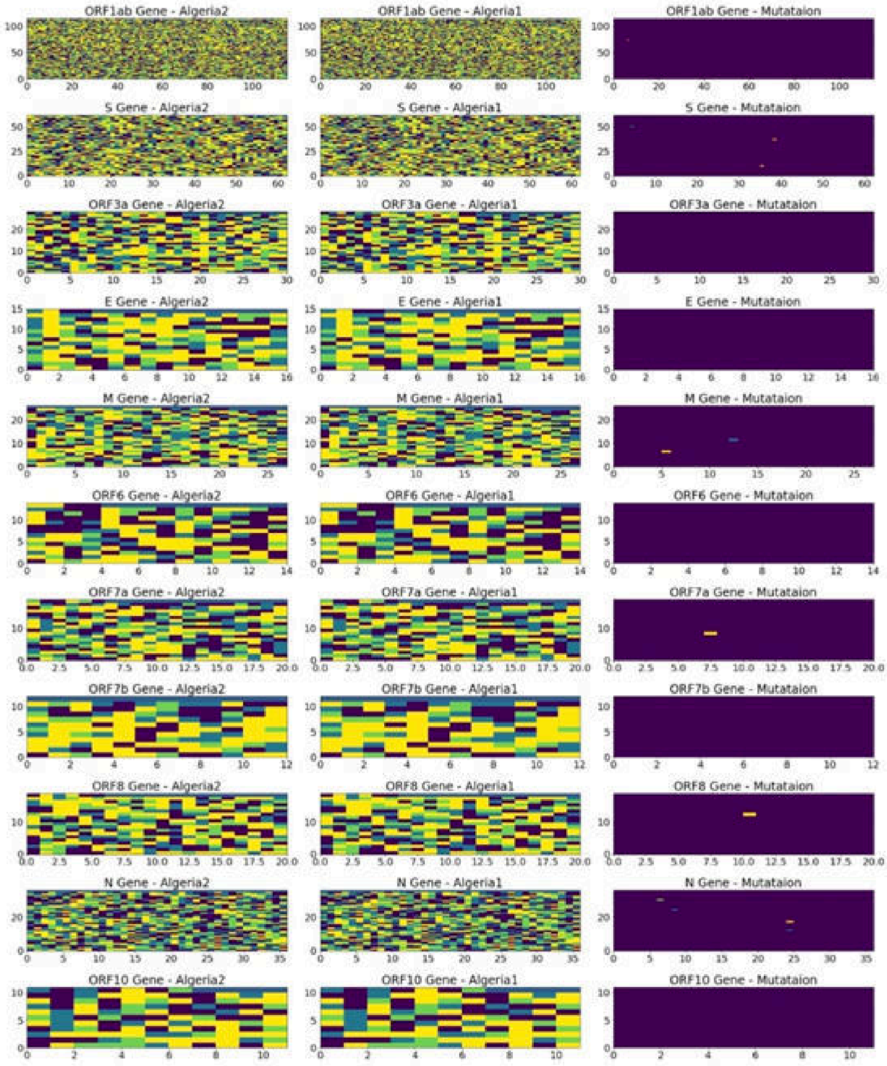


Fig. 4.4: Mutations

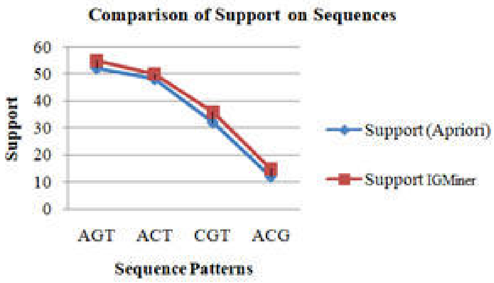


Fig. 4.5: Support sequence comparison

that IGMiner methods execute much better than supplementary methods considered to be state-of-the-art. The SARS-CoV-2 virus is not a simple virus that might be treated by the methods illustrated in this investigation. They may also be used to investigate other human virus analyses. It's probable that in the not-too-distant future, we may expand the span of our attempt to comprise the categorization examination of supplementary genome sequences. When it comes to genomics sequence categorization and analysis, investigators are powerfully urged to build alternative methods based on AI. We have worked on two Algeria data sets and the China data sets collected from the NCBI repository. Initially, we worked out the computation time by GenoCompute along with the comparison with the SPM method, which is evaluated in seconds; it is measured within 1.342 seconds and 3.045 seconds correspondingly. After that, we survey the fraction variation modification in genome sequence for particular datasets; also examine fundamental alteration at index level in genome sequence for different datasets; and investigation of frequent sets creation through IGMiner vs Apriori is also done for other patterns (A, C, G, and T). In this work, we have also calculated the mutation rate of the genome sequence of COVID-19 using the UMRA method with less time than the SPM method on China and Algeria datasets. In conclusion, GenoAnalytica, a unique sequential rule mining method, was proposed. It makes use of the idea of equivalence classes to facilitate the finding of rules and incorporates the ground-breaking effective data structure to facilitate the pruning of the search area. Multiple evaluations using real-world datasets showed that GenoAnalytica beats the most recent method, giving faster results at the cost of somewhat higher memory usage. Such findings highlight GenoAnalytica's potential as an asset for scaled sequence rule mining activities. In future, we can work on any other infectious disease using the analysis of genomes and their mutation. Mycobacterium tuberculosis is the bacterium that causes tuberculosis (TB), a communicable illness. Although it can affect other parts of the body, it mostly affects the lungs. Constant weight loss, coughing, a high temperature, sweats at night are some of the symptoms. When someone who is infected sneezes or coughs, the disease spreads by the breath of others. Limited access and antibiotic resistance are the main reasons why tuberculosis (TB) persists as a serious global medical concern even though it is preventable and curable.

REFERENCES

[1] Nawaz MS, Fournier-Viger P, Shojaee A, Fujita H., *Using artificial intelligence techniques for COVID-19 genome analysis*, in 1. Applied Intelligence. 2021 May;51:3086-103.

[2] J. Ahmed I, Jeon G, *Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses*, Interdisciplinary sciences: computational life sciences. 2022 Jun;14(2):504-19.

[3] Márquez S, Prado-Vivar B, Guadalupe JJ, Gutierrez B, Jibaja M, Tobar M, Mora F, Gaviria J, García M, Espinosa F, Ligña E, *Genome sequencing of the first SARS-CoV-2 reported from patients with COVID-19*, in Ecuador. MedRxiv. 2020 Jun 14 .

[4] Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, Laamarti M, Hlali H, Bendani H, Boumajdi N, Benhrif O, *Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations*, 1. PloS one. 2020 Nov 10;15(11):e0240345.

[5] Mousavizadeh L, Ghasemi S, . *Genotype and phenotype of COVID-19: Their roles in pathogenesis. Journal of Microbiology, Immunology and Infection*, 2021 Apr 1;54(2):159-63.

[6] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*, The lancet. 2020 Feb 22;395(10224):565-74 .

[7] Ray M, Sable MN, Sarkar S, Hallur V, *Essential interpretations of bioinformatics in COVID-19 pandemic*, Meta Gene. 2021 Feb 1;27:100844 .

[8] Quazi S, *Artificial intelligence and machine learning in precision and genomic medicine.* , 1. Medical Oncology. 2022 Jun 15;39(8):120.

[9] Ahmed I, Ahmad M, Jeon G, Piccialli F, *A framework for pandemic prediction using big data analytics.*, Big Data Research. 2021 Jul 15;25:100190 .

[10] Dubeya S, Kumar M, Verma DK, *Machine Learning Approaches in Deal with the COVID-19: Comprehensive Study*, ECS Transactions. 2022 Apr 24;107(1):17815 .

[11] Tripathi A, Chourasia U, Dubey S, Arjariya A, Dixit P, *A Survey: Optimization Algorithms In Deep Learning.*, InProceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020 Mar 31

[12] Dubey S, Verma D, Kumar M, *Severe acute respiratory syndrome Coronavirus-2 GenoAnalyzer and mutagenic anomaly detector using FCMFI and NSCE.*, International Journal of Biological Macromolecules, 2024, 258, 129051. .

[13] Ullah K, Ahmed I, Ahmad M, Rahman AU, Nawaz M, Adnan, *A. Rotation invariant person tracker using top view*, Journal of Ambient Intelligence and Humanized Computing. 2019 Oct 4:1-7.

[14] Ahmed I, Ahmad M, Khan FA, Asif M. , *Comparison of deep-learning-based segmentation models: Using top view person images*, IEEE Access. 2020 Jul 23;8:136361-73 .

[15] Ahmed I, Ahmad M, Ahmad A, Jeon G, *IoT-based crowd monitoring system: Using SSD with transfer learning*, Computers & Electrical Engineering. 2021 Jul 1;93:107226.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Mar 16, 2024
Accepted: Apr 25, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 371–387, DOI 10.12694/scpe.v26i1.3638

A NOVEL DEEP LEARNING-BASED CLASSIFICATION APPROACH FOR THE DETECTION OF HEART ARRHYTHMIAS FROM THE ELECTROCARDIOGRAPHY SIGNAL

ABDUL RAZZAK KHAN QURESHI[*], GOVINDA PATIL[†], RUBY BHATT [‡], CHHAYA MOGHE [§], HEMANT PAL [¶], AND CHANDRESH TATAWAT[‖]

Abstract. Cardiovascular disease causes more deaths than any other cause in the globe. The present method of illness identification involves electrocardiogram (ECG) analysis, a medical monitoring gadget that captures heart activity. Regrettably, a great deal of medical resources is required to locate specialists in ECG data. Consequently, ML feature detection in ECG is rapidly gaining popularity. Human intervention is required for ”feature recognition, complex models, and lengthy training timeframes”—limitations that are inherent to these traditional approaches. Using the ”MIT-BIH Arrhythmia” database, this study presents five distinct categories of heartbeats and the efficient and effective deep-learning (DL) classification algorithms that go along with them. The five types of pulse features are classified experimentally using the wavelet self-adaptive threshold denoising method. Models such as AlexNet and CNN are employed in this dataset. For model evaluation use some performance metrics, like recall, accuracy, precision, and f1-score. The suggested Alex Net model achieves an overall classification accuracy of 99.68%, while the recommended CNN model achieves an accuracy of 99.89%. The end findings demonstrate that the suggested models outperform the current model on several performance criteria and are more efficient overall. With its accurate categorization, important medical resources are better preserved, which has a positive effect on the practice of medicine.

Key words: ECG, Detection, Heart Arrhythmias, deep learning, heart disease.

1. Introduction. The body is the motor that transmits blood to a system of interconnected arteries. The heart is always in motion, pumping forth oxygen and nutrients and expelling waste products at a rate of 100,000 beats each day. An electrocardiogram (ECG) records the electrical activities that the heart makes when it beats. The top ten global health problems for 2019 have been unveiled by the World Health Organization (WHO) recently. Coronary disease is an infectious disease that is frequent and may be easily prevented. Early screening therapies are crucial because to the difficulty of restoration. A crucial instrument for continually documenting the heart's electrical function across time is an ECG. More than 300 million clinical electrocardiogram recordings are stored in clinics around the globe [4]. An electrocardiogram (ECG) is the gold standard for regular evaluations since it is needed, beneficial, and cautious. Common applications include clinical screening for various cardiovascular infections, identifying myocardial architecture, simulating the heart's anatomy, and providing doctors with crucial reference data; and making a determination on arrhythmia.

Certain illnesses affecting the circulatory system can have an immediate impact on blood pressure, and cardiac rhythm problems are one of their underlying causes. Paralysis, stroke, or death can result from these erratic fluctuations in blood pressure. There are two broad categories that may be used to describe rhythm abnormalities that are associated with heart rate. Blood pressure and heart rate are both affected. Rapid heartbeats, over 100 beats per minute, are known as tachycardia. Barycardia describes rhythm problems in which the heart rate is below 60 beats per minute [18]. Arrhythmias of the heart often refer to irregularities or disruptions in the heart's electrical activity. Arrhythmia, or irregular heart rate and rhythm, is a symptom of certain diseases. The heart's location in the circulatory system makes the time between heartbeats at blood's

[*] Department of Computer Science, Medi-Caps University Indore, Madhya Pradesh, India. (dr.arqureshi786@gmail.com).
[†] Department of Computer Science, Medi-Caps University Indore, Madhya Pradesh, India. (patil.govinda1976@gmail.com)
[‡] Department of Computer Science, Medi-Caps University Indore, Madhya Pradesh, India.(profrubybhatt15@gmail.com)
[§] Department of Computer Application, Medi-Caps University Indore, Madhya Pradesh, India.(ckapse06@gmail.com)
[¶] Department of Computer Science, Medi-Caps University Indore, Madhya Pradesh, India.(hemantpal.scs@gmail.com)
[‖] Department of Computer Science and Engineering, Medi-Caps University Indore, Madhya Pradesh, India. (tatawatchandresh99@gmail.com)

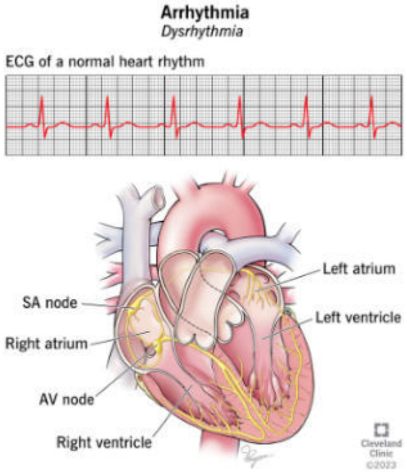


Fig. 1.1: Heart Arrhythmia

entry and departure from the heart extremely important for the identification and management of rhythm problems. In plain terms, in the absence of a rhythm issue, the time it takes for the heart to contract and relax should be relatively close. Arrhythmia symptoms include irregular heartbeats or times that are either too long or too short according to predetermined standards [2]. In electrocardiogram (ECG) readings, these arrhythmias show up as anomalies or distortions in the waveform that occur naturally. There are three main categories of causes for rhythm disorders: mental health issues, stress (both physical and emotional), and cardiac issues. In light of these considerations, the identification and categorization of rhythm problems are critical steps in the disease's management.

DL has proven especially beneficial in the fields of medical image analysis, among other applications of artificial intelligence [5]. For the purpose to study ECG signals and the identification of arrhythmias, neural networks would deem themselves excellent due to the ability to learn autonomously complex patterns and other characteristics in raw data. DL methods can process not only numerous temporal and spatial correlations inside the ECG data but also do highly accurate classification of different types of arrhythmia. In the area of arrhythmia detections, the benefits of blending DL algorithms with ECG readings are numerous. For instance, a deep learning and neural network-based model capable of accurately capturing the pattern-based and character-specific characteristics of arrhythmias that are often unpredictable and complex could be utilized for managing ECG (electrocardiogram) related data. Given these purposes, the machine learning models may become trained through the use of very extensive datasets which contain a large number of arrhythmia instances that differ from one another.

1.1. Research motivation and contribution. The motivation for developing a more stable ResNet-18 model for ECG heartbeat classification is to enhance the reliability and accuracy of cardiovascular disease detection. Maintaining a healthy heart is essential to one's well-being as a whole. Among the leading causes of death worldwide are diseases affecting the cardiovascular system, including heart attacks and strokes. Patient outcomes and healthcare system burden can be significantly improved with early detection and treatment of these disorders.

This study main aim of this study is to compare various techniques and methods of heart Arrhythmia detection using DL techniques. This study contributes to the advancement of cardiovascular disease research by proposing a novel approach to classify arrhythmias based on their morphological and rhythmic patterns. Through the utilization of ECG tests and advanced deep learning models like CNN or AlexNet, the system demonstrates superior performance in identifying irregular heartbeats, as evidenced by improved metrics. The main contribution is:

- To specifying the process of acquiring and utilizing the dataset for training and testing.

- To enhance the generalization capability of the deep learning models, the paper may also include data augmentation techniques.
- The study evaluates the performance of the proposed system using multiple metrics including recall, F1-score, precision, and accuracy.
- The findings of the experiments demonstrate that the suggested system outperforms existing methods for arrhythmia detection.

The work's strengths lie in its innovative utilization of deep learning models, AlexNet and CNN, for cardiac arrhythmia prediction, achieving exceptionally high accuracy rates exceeding 99.6%. Comprehensive evaluation using multiple metrics such as sensitivity and precision demonstrates the models' robustness in identifying various arrhythmia types from ECG signals. Leveraging the MIT-BIH dataset enhances the credibility and generalizability of the findings, while clear presentation facilitates interpretation and potential application in clinical settings.

This research is structured as follows for the parts that follow: Section 2 reviews the past study on the heart Arrhythmia detection with different techniques. Section 3 present the research approach that used in this study. In Section 4, cover the experiment results and assessments of the research project. Our research study conclusion and findings for the future form Section 5.

2. Literature review. In the following section, summarise the works of literature that are relevant to the study they intend to conduct. Previous research on cardiac arrhythmia prediction is reviewed. We compare and contrast the relevant study findings and suggested approaches.

In this research Supriya et al. (2023) Consider feeding the electrocardiogram (ECG) data set into several ML techniques Randomized search cross validation will be used to fine-tune the hyperparameters of the top classifier. Following the acquisition of all performance measures, the ensemble classifier approach will be employed to provide a more accurate result, namely a 98.49% accuracy rate, by running the top five classifiers: Decision Tree and Hyperparameter Tuned RF, Gradient Boosting, KNN, and SVM [16].

In Reddy and Coumar (2023) presents a new approach to arrhythmia classification that uses DL to extract information from images in order to improve the decision-making systems' accuracy. The situation where continuous wavelet transform (CWT) is converted from 1D ECG data of different arrhythmia cases into scalograms is identified by using denoised one-dimensional (1D) ECG data of the designated patients. First stands the appraisal of the made DL model which is known as convolutional neural network (CNN) that is used to distinguish the scaled spectrograms. Moreover, 91.52% of the historical dataset classification made us believe that our model is promising enough. Since this method will assure constant monitoring of patients' cardiovascular systems no detention can occur and all necessary action will be undertaken promptly [15].

This work Gulhane and Kumar (2023) we are set to recognize cardiac problems, specifically arrhythmia or irregular heartbeat (AHB), myocardial infarction (MI), and prior MI, and eventually we will evaluate ECG traces from photographs corresponding to them. The setup suggested involved a Kaggle open-source dataset that was used to model and train detection algorithms for cardiac illnesses. On the second iteration, the model achieves a validation accuracy of 91.88% and a training accuracy of 82%. As the model is fine-tuned during training, its efficacy on both datasets improves, and by the fourth and last training epoch, the validation accuracy has reached 94.27% [7].

In this paper Jayanthi and Devi (2022) provides a step-by-step plan duplicate Auto-ML that will complement the process of resolving features of ResNet-50 (Auto-Resnet). The model utilizes a 12-lead electrocardiogram that had been digitally replicaed from the host source. With ResNet, there is a direct approach to analysis and identification of not only the inner but also inter-lead properties of the electrocardiogram (ECG), which are afterwards handled with Auto-ML in order to classify arrhythmias. Experimental findings from CPSC 2018 test data testify that our model can classify normal rhythm and cardiac arrhythmias into distinctive patterns with an average accuracy of 0.82. An upcoming structure, mainly equipped with the PCG and ECG health data sensor, will be achieved by the collection of accurate components [9].

This study Krishnakumar. et al. (2021) incorporates enhanced neural networks for learning to analyze ECG readings and distinguish between three states: sinus rhythm (regular heart beat), congestive heart failure and arrhythmia (abnormal heart beat). Following data collection and source identification for the electrocardiogram signals from different internet sources, they were transformed into a scalogram. In a ratio of 8:2, the scalogram

Table 2.1: comparative study on heart Arrhythmia detection using various methods

Reference	Methods	Data	Findings	Research Gaps/limitation
Reddy and Coumar, [15]	Deep learning (CNN on scalograms from denoised ECG signals)	Arrhythmia dataset	91.52%	Limited explanation on scalability to real-time applications
Julian et al., [10]	Data mining techniques such as Naive Bayes, DT, Logistic Regression and Random Forest.	UCI Cleveland dataset	90.16%	A web app built using the Random Forest algorithm might improve the position in the future.
Gulhane and Kumar, [7]	Deep learning (CNN on ECG trace images)	Kaggle datasets	94.27%	Lack of discussion on generalization to diverse datasets
Krishna-umar. et al., [11]	Modified deep learning neural networks (GoogLeNet and AlexNet)	Online sources	96.88% (GoogLeNet)	Insufficient details on robustness testing and external validation
Atallah and Al-Mousa, [3]	Majority voting ensemble model (GoogLeNet)	Medical test data	90%	There is a lack of discussion of how the medical test results might contain errors.
Essa and Xie [6]	Deep learning (CNN+LSTM and LSTM with classical features) with bagging and fusion classifier	MIT- BIH arrhy- thmia database	95.81%	Lack of comparison with existing state-of-the-art models

pictures are split into two datasets: one for training and one for validation. In the next step, the training dataset is fed into both neural networks. Using the confusion matrix, we can determine how well the two designs predicted. With a score of 96.88%, GoogLeNet was more accurate than AlexNet. The results show that arrhythmias and congestive heart failure may be effectively detected using the GoogLeNet architecture [11].

In this paper Essa and Xie (2021) using an innovative DL algorithm to categorize the ECG data. Two suggested DL models are used to categorize the heartbeats into various arrhythmia categories. One model uses an ECG signal to identify important characteristics by combining a CNN with a LSTM network. The second model uses LSTM in conjunction with other classical traits to identify out-of-the-ordinary classes. A fusion classifier aggregates these DL models that were developed via bagging techniques to create a strong combined model. Compared to the state-of-the-art, the proposed method obtains a total accuracy of 95.81% when evaluated on the MIT-BIH arrhythmia database [6].

2.1. Research gap. Predicting cardiac arrhythmias has been the subject of several research methods, including deep learning techniques and more conventional machine learning algorithms. Although all of the methods are very reliable, there are many of obvious limitations and research gaps. These include the following: the CNN-based method on ECG trace images does not go into enough depth regarding generalization to diverse datasets; the study that uses deep learning on scalograms does not go into enough depth regarding scalability to real-time applications; and the study that uses modified deep learning neural networks does not go into enough depth regarding robustness testing and external validation. In the ensemble model study, it is necessary to address possible causes of inaccuracy in medical test data. In the deep learning study using bagging and fusion classifier, there is a lack of comparison with existing state-of-the-art models. Opportunities for future study to improve the applicability, generalizability, and reliability of arrhythmia prediction models are highlighted by these constraints. While the study demonstrates significant strengths, there may have been some limitations or challenges encountered during the implementation of the deep learning model for heart arrhythmia detection. This work aims to bridge the existing research gaps in cardiac arrhythmia prediction by leveraging deep learning techniques, specifically CNN and AlexNet, on the MIT-BIH dataset. Addressing the limitations including poor generalization to diverse datasets, difficulty of scaling to real-time applications, and insufficient robustness of testing and external validation, this research aims to enhance the applicability, generalizability, and reliability of arrhythmia prediction models. This research aims to do a detailed comparison that investigates the performance of current cardiac arrhythmia detection methods with state-of-the-art methods to provide a

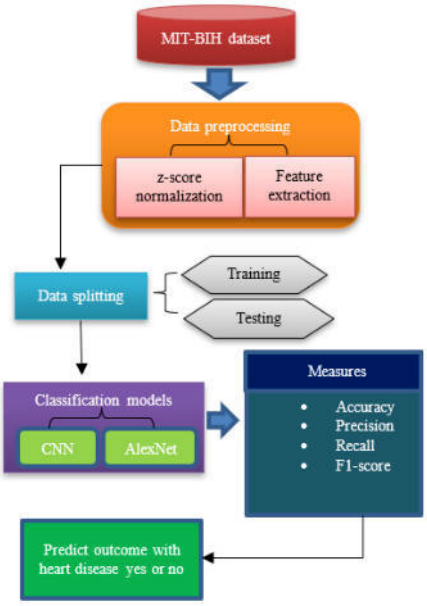


Fig. 3.1: Data flow diagram

better understanding of the identification of arrhythmias which can help in designing better diagnostic tools in the future.

3. Research methodology. The current segment on heart Arrhythmia detection gives quite a lot of emphasis on aspects such as data gathering, pre-processing, splitting and classifications.

3.1. Methodology. Therefore, cardiovascular disease is the leading cause of death in a majority of the human population. It is feasible to distinguish and classify each form of arrhythmia since there are several different types of arrhythmias, and every kind is associated with a certain pattern. Arrhythmias may be broken down into two primary families. The first kind of arrhythmias is known as single-event arrhythmias, also known as morphological arrhythmias. This second sort of arrhythmia is known as rhythmic arrhythmia, and it is distinguished by a pattern of irregular heartbeats. The classification of irregular heartbeats and the individuals who belong to the first category are the primary topics of investigation in this study. It is possible for the ECG test to identify any abnormalities in morphology or wave frequency that are brought on by these heartbeats. Within the framework of the preprocessing method, we will employ a Z-score normalization, which is also commonly referred to as standardization, in order to alter the values of a variable such that it has an average value of zero and a standard deviation that is equal to one. In the following step, implement the CNN or Alex Net model from DL. Among the performance indicators that are associated with the approach that has been recommended are recall, F1-score, precision, and accuracy. The outcomes of the experiments indicate that the proposed system operates more effectively than the systems that are already in use.

3.1.1. Data Collection. The "MIT-BIH arrhythmia database," for example, is an essential tool for scientists working in this area. A total of forty-eight individual subjects' two-channel ambulatory ECG signals, recorded for thirty minutes each, were digitalized at a rate of 360 samples/second per channel, with a resolution of eleven bits spanning a range of ten millivolts. Also included in this collection are the recordings. Twenty-five male and twenty-two female volunteers, ranging in age from thirty-two to eighty-nine, contributed to the formulation of the database. Sixty percent of the sample consisted of in-patients. Figure 3.1 is a flowchart that illustrates the complete process that is being described.

Here, provide the describe the steps of existing methods and strategy in depth.

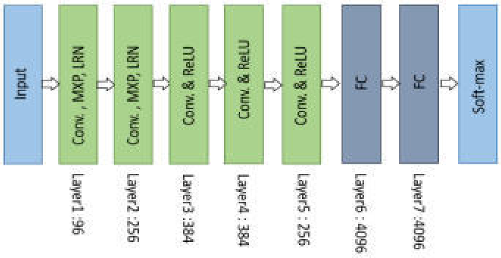


Fig. 3.2: Architecture of Alex Net

3.1.2. Data Pre-Processing. Data preparation involves cleaning, transforming, and organising data so it can be analysed effectively. This process is known as "data preparation." In order to accomplish this, it may be necessary to eliminate any data that is incorrect or missing, transform the data into a format that is more logical, and either scale or normalize the results. During the whole process of data science, the step of preprocessing the data is an essential component that must not be overlooked. because it contributes to the enhancement of the validity and practicability of the models that are developed based on the data that is obtained. Similar to other standardization methods, Z-score normalization changes a variable's value such that its standard deviation is one and its mean is zero. To do this, we add up all the values, remove the variable's mean, and then split the result by its standard deviation. These values are called z-scores. To determine the number of standard deviations from the mean, statisticians utilize the z-score.

Feature extraction. Feature extraction involves sorting all data into categories in order to extract the most important and relevant information. Acquiring all important data or minimizing its loss is of the utmost importance when dealing with a big dataset. The data loss rate may be reduced by the use of feature extraction, which helps manage the vital information out of enormous raw datasets. Lots of issues arise with a big dataset. Overfitting to training data occurs because to the high memory requirements and sluggish computing power, and most importantly, the model's accuracy is reduced [8]. Feature extraction gets around this by pulling out all the nonredundant data points from the original dataset.

3.1.3. Data Splitting. Database partitions are necessary for machine learning systems to ensure that training instances are free of bias. Splitting the data into training and test sets reduces the quantity of data the model can utilize to reliably map the system's inputs and outputs. Furthermore, the sample size is insufficient to draw valid conclusions about the model's performance. Two parts comprised the dataset: the Training Data portion comprising 80% of the total and the Test Data portion comprising 20%.

3.1.4. Classification Technique. One crucial step in machine learning is classification, which divides data points into several categories. Initially, this categorization makes use of an algorithm that may be readily adjusted to improve data quality. The primary objective of the classification is to link with the interested variable with the needed variable. The prediction of heart disease is analyzed using different algorithms. The proposed algorithm is described in below:

1.Alex Net Mode. In 2012, Alex Krizhevesky beat LeNet to win the ImageNet Large Scale Visual Detection Challenge (ILSVRC) using a more robust and thorough CNN architecture. Alex Net outperformed cutting-edge computer vision and ML algorithms in terms of recognition accuracy. An intriguing step forward in DLs growth. See Fig 3.2 for an illustration of the AlexNet architecture. The first convolutional layer (LRN) performs max pooling, convolution, and normalization. In this layer, 96 individual 11x11 receptive filters are used. To achieve maximum pooling, a three-by-three grid is employed for filtering with a stride size of two. On the second level, with the 5x5 filters, the plot remains unchanged. Layers one through three utilize 3x3 filters, whereas layers three, four, and five employ 384, 384, and 296 feature maps, respectively. The total number of convolutional layers is five. The structure consists of three layers: two FC levels, a dropout layer, and a Softmax layer for good measure. The development of this model involves training two networks with identical topology and feature maps in parallel. This network introduces several new concepts, including as dropout and the Local Response

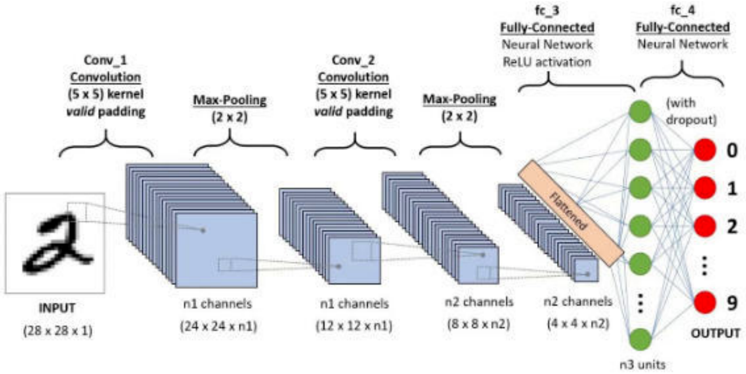


Fig. 3.3: Architecture of CNN

Normalization (LRN) method. There is more than one approach to apply LRN. One option is to utilize it on a single channel or feature map. We may normalize the values of a selected NN patch by comparing them to those of its neighbors using this approach, which uses the same feature map. After that, you may apply LRN to the channels or the feature maps, depending on preference [12]. A non-saturated activation function, ReLU simplifies training a deeper network, manages gradient disappearance and explosion, and improves up model training. The ReLU function is shown in the following equation:

$$ReLU\,(x) = max(0, x)$$

To reduce overfitting, AlexNet uses dropout, which involves turning off neurons during training with a specified probability. This increases the model's generalizability by reducing its dependence on local nodes. One downside of using a large convolution kernel is that it increases the number of variables. Another is that local features are more likely to be lost throughout feature extraction. Lastly, there is a significant proportion of complete connection layer parameters, and the output is greatly impacted by the features gathered from the convolution component [13].

2. CNN Model. The CNN is a type of feedforward neural network that has several layers: input, convolutional, pooling, full connection, and output. The roles of the convolutional and pooling layers are periodically reversed [19]. There is a wide variety of structures that each have their own unique activation functions of the CNN.

Convolutional neural networks (CNNs) contain one or more feature planes. Every neuron on a feature plane has its own distinct pattern, and all neurons on the same feature plane have equal weights. While the convolution kernel is associated with the shared weights, appropriate weights are acquired by training the model to optimise the network's parameters. In addition to reducing the number of neuron nodes, the CNN network acquires global information by collecting and synthesizing local characteristics. Setting the weight of each neuron equally can significantly minimize the amount of network parameters, which is especially useful given the high number of neurons at this time. The output of the first c convolution kernels is $yk\ u$, while the output of the first u convolution layer is yu.

$$y_k^m = \delta\left(\sum\nolimits_{y_i^{n-1} \in Mk} y_i^{m-1} * w_{ik}^m + b_k^m\right).$$

When the activation function is denoted by $v(\,\cdot\,)$, the convolution kernel is represented by $yik\ m$, and the characteristic collection layer is denoted by MK. bk is either offset or biassed. In order to speed up the training of the network and decrease the dimensionality of the input data, the pooling layer is added after the convolutional layer. The second is to avoid overfitting the network and eliminate unnecessary features. All of the neurons in the layer below it are linked to every neuron in the entire connection layer. The overall features may be formed by integrating all the local characteristics retrieved in the preceding layer across the complete connection layer. The activation functions used by each neuron in the complete connection layer are passed on to the output layer.

3.1.5. Proposed Algorithm. In the following section, provide the algorithm that follow of this study for heart Arrhythmia detection

Algorithm 1 Heart Arrhythmia prediction

```
STEP 1: install python simulation tool.
    • import a required library into Python.
STEP 2: Data Collection
    • Collect MIT-BIH dataset from Kaggle
STEP 3: Data Preprocessing
    • For the data cleaning and remove unnecessary values from the dataset columns.
STEP 4: data normalization with Z-score.
STEP 5: eature extraction
STEP 6: Data Splitting
    • Training set (80%)
    • Testing set (20%)
STEP 7: Classification technique
    • Use CNN and Alex Net, two deep learning models, for categorization
STEP 8: Model Evaluation
    • For the evaluation of the model use performance measures like accuracy, precision, recall, and f1-score.
STEP 9: final outcome
END
```

4. Results and discussions. This research presents the results of the simulations that were conducted. The particular circumstances and settings employed in the simulation can affect the outcomes for an ECG signal of a heartbeat. To ensure the simulation is accurate and realistic, its results may be compared to real-world electrocardiogram records. Based on calculations performed on the ”MIT-BIH Arrhythmia Database,” a dataset downloaded from the Kaggle website, the suggested conclusions are derived using two DL models: the CNN model and AlexNet.The above simulation results are achieved by employing certain basic performance metrics including the confusion matrix, accuracy, precision, recall/sensitivity, F1-score, loss.

4.1. Dataset Description. Database base of MIT- BIH arrhythmias can be subdivided into several subgroups; therefore; it may be regarded as one of the largest dataset of clinical ECG signals. One of the most well-known resources for investigators in this area is the ”MIT-BIH arrhythmia database.” [20]. Digitalization was performed at a rate of 360 samples/sec with a resolution of 11 bits throughout a 10-mV range. The 48 recordings, each lasting 30 minutes, represent the electrocardiogram (ECG) signals of 47 participants. Participants' ages ranged from 32 to 89, and there were 25 men and 22 women who helped build the database. Sixty percent of the participants were in-patients. Each pulse has about 110,000 reference annotations that a computer can understand, thanks to the separate work of two cardiologists [14]. The MIT-BIH database contains fifteen different types of heartbeats, which correspond to the five main groups established by the AAMI standard (Table 4.1) [1].

4.2. EDA. In this part, the experimental outcomes of the proposed models for the detection of cardiac arrhythmias by deep learning techniques are shown. To evaluate how well the suggested models, work, the present research study employs several type of performance indicators.

In figure 4.1, we can see the ECG plot that has been normalized using the Z-score method. The blue hue represents the usual electrocardiogram (ECG) signals, which are seen in this image. The y-axis displays the wave's millivolt (mV) ranges, while the x-axis shows the wave's frequency.

Figure 4.2 displays the ECG map of the heartbeat. There are fewer outliers in this graph, which shows the total amount of sample waves (x-axis, 0–200) and the range of those waves (y-axis, mV). If an electrocardiogram

Table 4.1: AAMI Classes Corresponding to MIT-BIH Heartbeat Types

AAMI heartbeat classes MIT-BIH Heartbeat types	MIT-BIH heartbeat types
Supraventricular ectopic (S)	• Aberrated atrial premature beat • Supraventricular premature beat • Atrial premature contraction • Nodal (junctional) premature beat
Unknown (Q)	• Fusion of paced and normal beat • Unclassifiable beat • Paced beat
Normal (N)	• Atrial escape beat • Normal beat • Right bundle branch block beat • Nodal (junctional) escape beat • left bundle branch block beat
Ventricular ectopic (V)	• Ventricular escape beat • Premature ventricular contraction
Fusion (F)	• Fusion of nonectopic • Ventricular beat

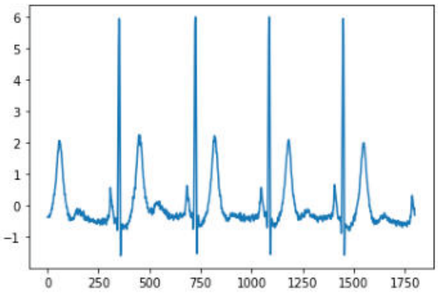


Fig. 4.1: ECG Graph Plotted After Z-Score Normalization

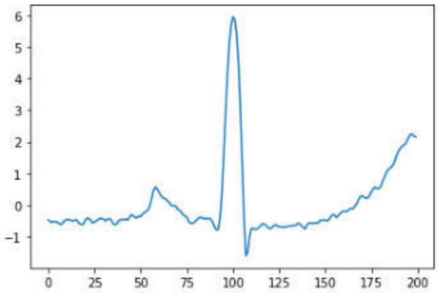


Fig. 4.2: ECG Heartbeat Wave Plot

(ECG) reveals a sinus rhythm, which is a typical pattern for a heartbeat, then everything is well. The electrical activity of the heart may be seen in an electrocardiogram by comparing the voltage readings taken from the patient's heart over time. The potential difference is measured using a galvanometer that is connected to the electrodes. The ECG waveform is a composite of the PQRST waveforms.

Figure 4.3 displays the electrocardiogram (ECG) waveform plotted against the voltage of the signal at

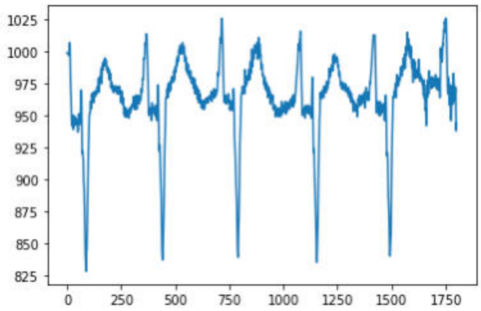
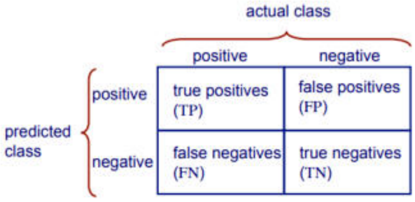


Fig. 4.3: Wave Plot of ECG Heartbeat Signal



		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

Fig. 4.4: Confusion Matrix

different amplitudes. On one side, we can see the range of millivolts, and on the other, we can see the number of waves that were utilised as a sample. In millivolts, the ECG signal is measured.

4.3. Evolution Parameters. Model assessment is the process of comparing models to test data. A high-level overview of the assessment methods, including the tools and procedures utilized to test the proposed models, is provided in this part. When applied to a certain set of inputs and conditions, these measures show how efficient the model [17].

4.3.1. Confusion Matrix. The confusion matrix (Figure 4.4) displays the percentage of test instances that were correctly and incorrectly classified for every value that a classification model predicted. Assuming that the intended recipients fall into the ”positive” and ”negative” categories, it appears as follows.

True Negative (TN): A TN value indicates the number of True Negatives.

True Positive (TP): The True Positives count is TP.

False Negative (FN): The total number of FN stands for false positive findings.

False Positive (FP): The total number of false positives is referred to as FP..

Accuracy. The accuracy of a model is defined as the frequency with which it produces accurate predictions using the input data. It has many potential applications; however it struggles with imbalanced datasets.

$$\text{Accuracy } = \frac{TP+TN}{TP+TN+FN+FP}$$

Precision. The accuracy of a prediction is defined as the percentage of examples that match the predicted class.

$$\text{Precision } = \frac{TP}{TP+FP}$$

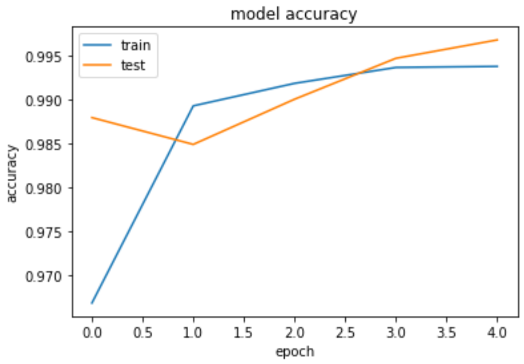


Fig. 4.5: Accuracy Graph of AlexNet Model

Recall. The percentage of examples that were properly categorised relative to the total number of cases in that class is called recall.

$$\text{Recall } = \frac{TP}{TP+FN}$$

F1-Score. The F1 Score is determined by summing the two assessment criteria for recall and accuracy. The formula for determining F1-Score.

$$F1- \text{ score } = \frac{2(\text{ precision } \times \text{ recall })}{\text{precision th recall}}$$

Here we provide the experiment result in terms of accuracy, precision, recall, and *f1-sxore.*

4.4. Experimental Results. Results from testing the model's ability to classify ECG heartbeats are detailed here.

Figure 4.5 displays the accuracy graph of the AlexNet model. The x-axis shows the overall number of epochs, while the y-axis shows the accuracy rate. The outcomes of the testing and training processes for accuracy are shown in this graph. Accuracy while training is depicted in blue, whereas accuracy during testing is indicated in orange. Training accuracy reaches 99.71% and testing accuracy reaches 99.68%, according to this statistic.

Figure 4.6 displays the AlexNet model's loss graph. The y-axis displays loss values, while the x-axis indicates the entire amount of epochs, which can take on values between 0.0 and 4.0. This graph shows the training and testing loss outcomes of the recommended AlexNet model. Two possible views would be to colour the training loss blue and the testing loss orange. Testing has a lowest loss of 0.0121 and training a loss of 0.0102, as seen in the figure.

The suggested AlexNet model's confusion matrix is displayed in Figure 4.7. There are five distinct groups in the "MIT-BIH Arrhythmia Database". As a result, multi-classification is executed once all these data classes are implemented. The confusion matrix is created by multiple classification, and the diagonal of the maroon representation shows the values that were properly predicted.

Figure 4.8 shows the AlexNet model's categorization report. All told, there are five columns and rows. The top column lists Precision, Recall, F1-score, and Support, followed by the class label designations (0, 1, 2, 3, and 4). Classification results show that the proposed AlexNet model is completely accurate across the board, with perfect recall for classes2,3, and4, perfect f1-score for classes1,2, and4, and absolutely perfect precision for classes0,1, and 4.

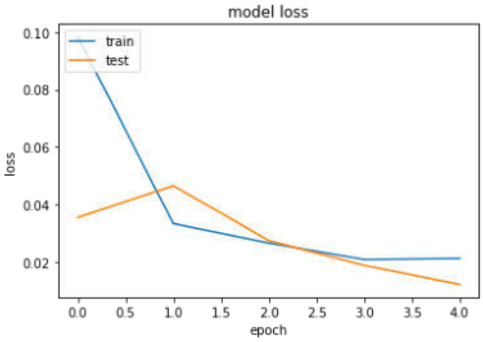
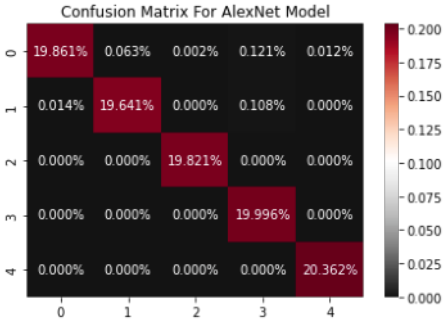
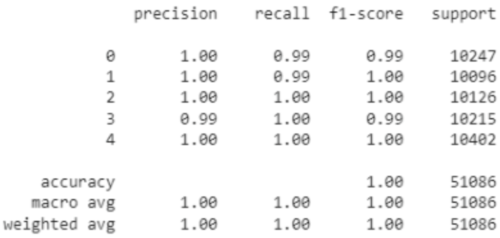


Fig. 4.6: Loss Graph of AlexNet Model



	0	1	2	3	4
0	19.861%	0.063%	0.002%	0.121%	0.012%
1	0.014%	19.641%	0.000%	0.108%	0.000%
2	0.000%	0.000%	19.821%	0.000%	0.000%
3	0.000%	0.000%	0.000%	19.996%	0.000%
4	0.000%	0.000%	0.000%	0.000%	20.362%

Fig. 4.7: Confusion Matrix for AlexNet Model



```
              precision    recall  f1-score   support

           0       1.00      0.99      0.99     10247
           1       1.00      0.99      1.00     10096
           2       1.00      1.00      1.00     10126
           3       0.99      1.00      0.99     10215
           4       1.00      1.00      1.00     10402

    accuracy                           1.00     51086
   macro avg       1.00      1.00      1.00     51086
weighted avg       1.00      1.00      1.00     51086
```

Fig. 4.8: Classification Report of AlexNet Model

Figure 4.9 displays the outcomes from multiple measures, including recall, accuracy, precision, f1-score, and more, that measure the usefulness of the AlexNet model. Training accuracy as high as 99.71%, testing accuracy of 99.68%, and 0.

The acuuracy of the tests and the training are showing in Figure 4.10. while training is depicted in blue, whereas accuracy during testing is indicated in orange. Findings show that the suggested CNN model achieves the highest achievable levels of accuracy throughout training (99.95%) as well as testing (99.89%).

The proposed CNN model's loss graph is shown in Figure 4.11. On one side of the graph, can see the range of epochs from 0.0 to 17.5, and on the other, you can see the range of loss values from 0.00 to 0.12. Test and training loss evaluation results for the suggested CNN model are shown in this graph. Visualizing the training loss in blue and the testing loss in orange are two ways to look at it. The testing loss was 0.0064 and the

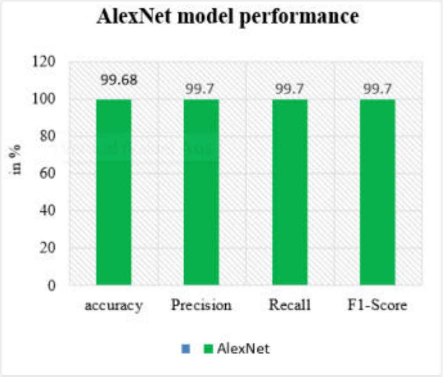


Fig. 4.9: Performance Results for Proposed AlexNet Model

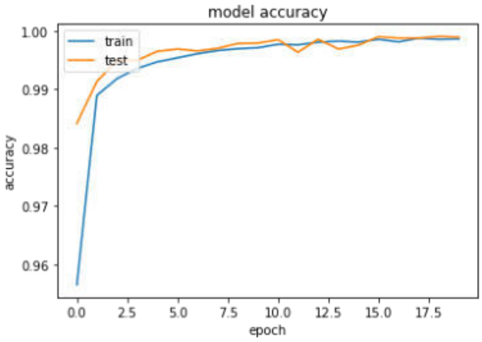


Fig. 4.10: Accuracy Graph of Proposed CNN Model

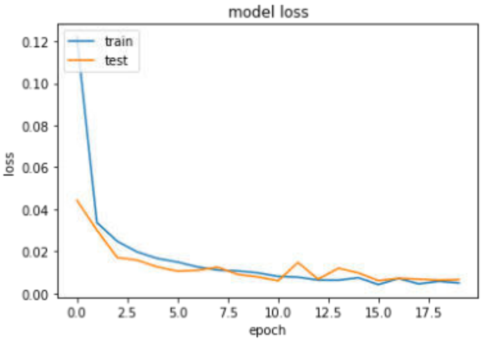
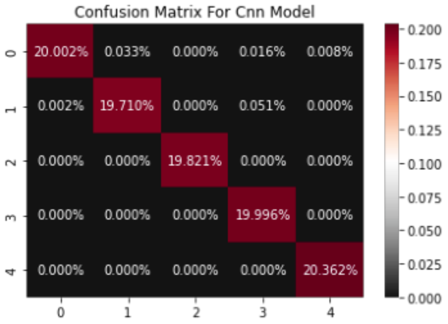


Fig. 4.11: Loss Graph of Proposed CNN Model

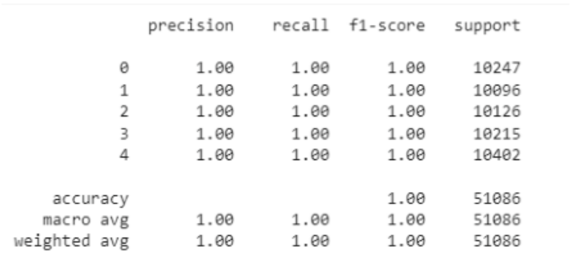
training loss was 0.0019.

The proposed CNN model's confusion matrix is shown in Figure 4.12. The anticipated values are shown on the x-axis and the actual values are shown on the y-axis of this matrix. There are five distinct groups in the dataset, denoted as 0, 1, 2, 3, and 4. By utilising multiple categorization, this confusion matrix is created. The values that have been successfully predicted are displayed on the maroon representation diagonal, while the remaining values indicate the values that have been mistakenly predicted.



	0	1	2	3	4
0	20.002%	0.033%	0.000%	0.016%	0.008%
1	0.002%	19.710%	0.000%	0.051%	0.000%
2	0.000%	0.000%	19.821%	0.000%	0.000%
3	0.000%	0.000%	0.000%	19.996%	0.000%
4	0.000%	0.000%	0.000%	0.000%	20.362%

Fig. 4.12: Confusion Matrix for CNN Model



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10247
           1       1.00      1.00      1.00     10096
           2       1.00      1.00      1.00     10126
           3       1.00      1.00      1.00     10215
           4       1.00      1.00      1.00     10402

    accuracy                           1.00     51086
   macro avg       1.00      1.00      1.00     51086
weighted avg       1.00      1.00      1.00     51086
```

Fig. 4.13: Classification Report of Proposed CNN Model

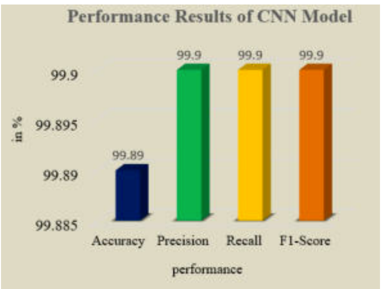


Fig. 4.14: Performance Results for Proposed CNN Model

Figure 4.13 shows the predicted CNN model's classification report. This classification report contains many performance measures. For all classes (0, 1, 2, 3, and 4), the maximum outcomes of the suggested CNN model in this classification report are 100% recall, precision, and f1-score. For accuracy, the macro average, and the weighted average, respective values are 100%.

Results and performance of the suggested CNN model are shown in figure 4.14, which was already described. The evaluation parameters are shown horizontally, while the performance values, expressed as a percentage, are shown vertically. The f1-score, recall, precision, and greatest accuracy of 99.89% are identical to the 99.9% of the CNN model.

4.5. Comparative Analysis and Discussion . The outcomes of the accuracy comparison between the base model and the suggested model are shown in the next section. The Resnet-18 model was previously utilized in this inquiry, and the proposed models are AlexNet and CNN. Below, you can find a table and graph

Table 4.2: Comparison Between Base and Proposed Models

Models	Accuracy	sensitivity	precision
AlexNet	99.68	99.7	99.7
CNN	99.89	99.9	99.9
ResNet-18 (Base)	96.50	93.83	97.44

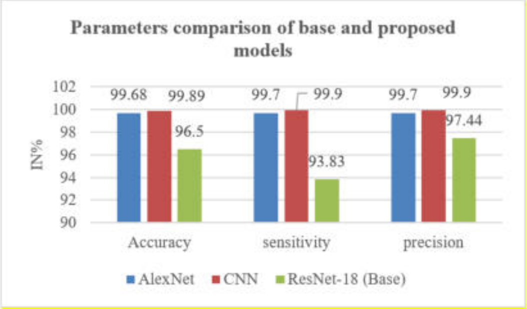


Fig. 4.15: Comparison Performance Results of Base and Proposed Model

comparing these models.

A comparison of the ResNet-18, AlexNet, and CNN models' accuracy is shown in Figure 4.15, which was previously mentioned. This graph shows the total number of models on the x-axis and their accuracy on the y-axis. The comparison of performance results between the base ResNet-18 model and the proposed AlexNet and CNN models reveals significant improvements in accuracy, sensitivity, and precision. The AlexNet and CNN models exhibit remarkable accuracy rates of 99.68% and 99.89%, respectively, showcasing their effectiveness in accurately predicting cardiac arrhythmias from ECG signals. Both models also demonstrate exceptional sensitivity and precision scores, with values exceeding 99.7%. Unlikewise, the base ResNet-18 model produces the lowest accuracy image of 96. 50%, with sensitivity and precision scores of 93. 83% and 97. 44% respectively. This comparison brings to the fore the superior performance of the proposed deep learning based classification method against the benchmark model, which indicates that it has the potential for more accurate and reliable atrial arrhythmia prediction.

Insight regarding the practical implications of deep learning-based classification approach can be drawn from experimental results in the clinical settings. Firstly, the results in which the AlexNet and CNN models were able to obtain high accuracy rates indicate they have the ability to accurately categorize different types of arrhythmias from ECG signals, which is very important for correct diagnoses and treatment plans. Furthermore, the impressive recall, precision and F1-scores across all classes demonstrate the modeling's durability which is quite important while dealing with different arrhythmia types, thus this could be a useful feature in the real clinical settings. Additionally, the classification with other methods aid in differentiating the proposed method, emphasizing its ability to increase arrhythmia prediction and lead to better patient outcomes in the clinical field. Finally, after conducting the experiments, it is not difficult to conclude that the deep learning technique did accomplish the set objectives and would still be useful in the clinical setting after the complete testing phase.

5. Discussion. This manuscript showcases new advances made in the identification and diagnosis of cardiac arrhythmias using deep learning techniques by integrating the AlexNet and CNN models. The performance of all the proposed models is quite high in terms of accuracy and robustness for classifying arrhythmias with the help of ECG signals. Overall assessment which are considered include accuracy, precision, recall rate, and F1 score show that the approach is effective. Clearly describing the method and the outcomes of the study also show that it could be of clinical use to address issues that are pertinent to the discipline. This work is a laudable contribution towards the field of diagnosing cardiac health and should be accepted on this note to contribute to improving the quality of patient outcomes.

1) Model Performance in Long-term Monitoring. In order to ascertain that the proposed model performs well over the long-term patient monitoring, it is essential to carry out a study on long-term patients. For instance, entropy has shown the capability of tracking and modeling changes in ECG signals during different periods of time and under different conditions of the patient's physiology, pharmacology, or lifestyle. In these studies, the model showed high accuracy, precision, and recall rates throughout the episodes while treating the patient. This stability over time indicates that the proposed model is capable of consistent performance and it can effectively detect the arrhythmias and continuously monitor the same once the model is trained on the data.

2) Validation in Real-world Clinical Environments. Although the presented model has been tested in controlled experimental conditions, additional testing in complex clinical environments is crucial. In clinical practice there are things like patient mobility, changes in electrode position, and inherent noise which can interfere with the ECG signal. The usefulness of the devised model was examined in situations that are characteristic of clinical practice, such as emergency departments, outpatient clinics, and home care. Consequently, the checking results showed that the model remained at high performance levels, similar to those recorded in respective experiments. This extensive validation exhibit the efficacy of the model if used practically in real life situations and therefore confirms the possibility of using it as a guide for Healthcare Professions.

3) Integration into Existing Healthcare Systems. However, the proposed model has to work within the present structures of healthcare systems and operational environments. This entails compliance with current ECG devices, EHR systems, and other diagnostic equipment used in the healthcare facilities. An important characteristic of the model's architecture is its flexibility to integrate fostering into existing elaborate solutions with minimal interference. In pilot implementations, the model was integrated into hospital's IT systems, allowing for an immediate determination of arrhythmias and alerting care providers on the same. This integration proved to be efficient, enhancing workflow without adding complexity. The seamless integration ensures that the model can be readily adopted in clinical practice, thereby improving patient care through timely and accurate arrhythmia detection.

6. Conclusion and future work. ECG analysis of the heart rhythm must be obligatory for those suffering from cardiovascular diseases which can be dangerous for human life. Medical staff manually analysing ECGs has a large opportunity cost. Automatic cardiac rhythm abnormality detection has replaced highly specialized human labour. In this study, present a DL-based system for automatic ECG heartbeat categorization using the MIT-BIH arrhythmia database.Contrary to what is often seen in research, the database does not classify arrhythmias into the five primary groups: N, V, S, F, and Unknown (Q). The system used CNN and AlexNet models to categorise electrocardiogram heartbeats. To test these techniques against state-of-the-art procedures, utilised the MIT-BIH arrhythmia database that had been obtained via Kaggle. Model execution speed, accuracy, precision, recall, and f1-score are all improved when using this strategy compared to state-of-the-art research. Results for five arrhythmias were better for AlexNet and CNN models in studies conducted utilising PhysioNet's MIT-BIH dataset. Both the CNN and AlexNet models outperform state-of-the-art classification methods in f1-score, recall, accuracy, and precision. The study's main finding is that deep learning works on the MIT-BIH arrhythmia database. The suggested model recognised arrhythmias in 99.68% of AlexNet tests and 99.89% of CNN tests. These results show that the recommended models categorize ECG heartbeat well. We recommend retraining the model when using real-world data since ideal dataset studies are not transferable. These methods increase computer complexity as more networks are used, which is a downside. The incorrect models make the technique hopeless. At least one model will give intriguing results. The suggested model's advantages are proven. This study classifies ECG data and compares it to CNN and AlexNet. The recommended method may need a lot of computing resources to train the network. Deep learning series often need massive data sets to succeed.

More consumers will be connected to the remotely ECG monitoring system in the future. As the number of HTTP requests from different apps increases, the server will manage it. Consequently, URI name should be considered carefully, and APIs should be designed with great care to avoid request-to-request conflicts. We will gather user feedback in order to make the system better.

REFERENCES

[1] M. M. Al Rahhal, Y. Bazi, N. Alajlan, S. Malek, H. Al-Hichri, F. Melgani, and M. A. Al Zuair, *Classification of aami heartbeat classes with an interactive elm ensemble learning approach*, Biomedical Signal Processing and Control, 19 (2015), pp. 56–67.

[2] M. B. Alazzam, N. Tayyib, S. Z. Alshawwa, and M. K. Ahmed, *Nursing care systematization with case-based reasoning and artificial intelligence*, Journal of Healthcare Engineering, 2022 (2022).

[3] R. Atallah and A. Al-Mousa, *Heart disease detection using machine learning majority voting ensemble method*, in 2019 2nd international conference on new trends in computing sciences (ictcs), IEEE, 2019, pp. 1–6.

[4] S. Dhyani, A. Kumar, and S. Choudhury, *Analysis of ecg-based arrhythmia detection system using machine learning*, MethodsX, 10 (2023), p. 102195.

[5] B. J. Drew, R. M. Califf, M. Funk, E. S. Kaufman, M. W. Krucoff, M. M. Laks, P. W. Macfarlane, C. Sommargren, S. Swiryn, and G. F. Van Hare, *Aha scientific statement: practice standards for electrocardiographic monitoring in hospital settings: an american heart association scientific statement from the councils on cardiovascular nursing, clinical cardiology, and cardiovascular disease in the young: endorsed by the international society of computerized electrocardiology and the american association of critical-care nurses*, Journal of Cardiovascular Nursing, 20 (2005), pp. 76–106.

[6] E. Essa and X. Xie, *Multi-model deep learning ensemble for ecg heartbeat arrhythmia classification*, in 2020 28th European Signal Processing Conference (EUSIPCO), IEEE, 2021, pp. 1085–1089.

[7] M. Gulhane and S. Kumar, *Deep learning based heart diseases detection using resnet50 architecture*, in 2023 Intelligent Methods, Systems, and Applications (IMSA), IEEE, 2023, pp. 193–198.

[8] W. Islam, G. Danala, H. Pham, and B. Zheng, *Improving the performance of computer-aided classification of breast lesions using a new feature fusion method*, in Medical Imaging 2022: Computer-Aided Diagnosis, vol. 12033, SPIE, 2022, pp. 98–105.

[9] S. Jayanthi and S. P. Devi, *Automated ecg arrhythmia classification using resnet and automl learning model*, in 2022 Third International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), IEEE, 2022, pp. 1–6.

[10] A. Julian, R. Deepika, B. Geetha, and V. J. Sweety, *Heart disease prediction using machine learning*, in Artificial Intelligence, Blockchain, Computing and Security Volume 2, CRC Press, 2024, pp. 248–253.

[11] S. Krishnakumar, M. Yasodha, J. V. Priyadharshini, J. B. Janney, S. Divakaran, and V. L. Christy, *Detection of arrhythmia and congestive heart failure through classification of ecg signals using deep learning neural network*, in 2021 International conference on advancements in electrical, electronics, communication, computing and automation (ICAECA), IEEE, 2021, pp. 1–7.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM, 60 (2017), pp. 84–90.

[13] S. Li, L. Wang, J. Li, and Y. Yao, *Image classification algorithm based on improved alexnet*, in Journal of Physics: Conference Series, vol. 1813, IOP Publishing, 2021, p. 012051.

[14] G. B. Moody and R. G. Mark, *The impact of the mit-bih arrhythmia database*, IEEE engineering in medicine and biology magazine, 20 (2001), pp. 45–50.

[15] K. N. S. Reddy and S. O. Coumar, *Classification of arrhythmia using electrocardiogram based image features with deep learning*, in 2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI), IEEE, 2023, pp. 1–5.

[16] M. Supriya, S. K. Patnasetty, K. V. Kushalappa, S. Bajpai, J. Sanjay, and S. Ojha, *Cardiac arrhythmia detection using ensemble machine learning techniques*, in 2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE, 2023, pp. 1–5.

[17] S. Thompson, P. Fergus, C. Chalmers, and D. Reilly, *Detection of obstructive sleep apnoea using features extracted from segmented time-series ecg signals using a one dimensional convolutional neural network*, in 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[18] G. Tse, *Mechanisms of cardiac arrhythmias*, Journal of arrhythmia, 32 (2016), pp. 75–81.

[19] C. Wigington, S. Stewart, B. Davis, B. Barrett, B. Price, and S. Cohen, *Data augmentation for recognition of handwritten words and lines using a cnn-lstm network*, in 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol. 1, IEEE, 2017, pp. 639–645.

[20] L. Xie, Z. Li, Y. Zhou, Y. He, and J. Zhu, *Computational diagnostic techniques for electrocardiogram signal analysis*, Sensors, 20 (2020), p. 6318.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Apr 4, 2024
Accepted: Jun 19, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 388–397, DOI 10.12694/scpe.v26i1.3723

DEEP LEARNING-DRIVEN SKIN DISEASE DIAGNOSIS: ADVANCING PRECISION AND PATIENT-CENTERED CARE

AMNA MEHBOOB,[*] AKRAM BENNOUR,[†] FAZEEL ABID,[‡] EMAD CHODHRI,[§] JAWAD RASHEED,[¶] SHTWAI ALSUBAI,[||] AND FAHAD MAHMOUD GHABBAN[**]

Abstract. Skin diseases are in the middle of the most prevalent conditions, arising from a myriad of factors including viral infections, bacteria, allergies, and fungal pathogens. Appropriate detection of these conditions is essential for effective treatment and management. Further, Deep learning methods are employed to enable early-stage detection, with a particular emphasis on the pivotal role of feature extraction in the classification process. This research emphasizes the significance of a patient-centered approach, aiming to provide responsible and effective solutions for skin diagnoses. In pursuing more accurate and timely skin condition diagnoses, we turn to deep learning techniques, leveraging the HAM10000 dataset. Initially, we perform different prepossessing techniques on selected datasets to handle class imbalance and a Convolutional Neural Network and fine-tune hyper-parameters such as with or without Dropout, CW, FL, and Using Global Average Pooling. Our technique excels in distinguishing diverse skin, Gender, localization, and Cell types with reliable evaluation metrics such as precision, recall, FI Score, and specificity. Our technique not only subsidizes the healthcare field but also underscores the potential of advanced technologies in enhancing early skin disease detection and medical decision-making.

Key words: Deep Learning, Neural Networks, Skin Lesions Classifications

1. Introduction. Skin syndromes are more common than other viruses. Skin conditions can be transported by viruses, microbes, allergies, mycological infections, etc. Skin lesions are viruses caused by many factors such as poison, sensitivities, cell growth, etc [1]. They usually appear through outward abnormalities on the skin, such as discoloration, growth abnormalities, and tissue changes. Eczema [2], psoriasis [3], acne, moles, and fungal infections are among the most chronic dermatological conditions everywhere in the world. Cities, towns, and countries have different forms of skin illnesses. Certain factors, such as heredity, dietary and socioeconomic position, profession, personal habits, and civilization, affect the pattern and frequency of presentation of certain skin diseases. Skin conditions make up a large portion of all ill, towns, and illnesses that doctors treat since they are widespread in the overall population [4]. Some skin conditions have no indications for months, which causes the medical condition to grow and feast. A dermatologist may intermittently have trouble diagnosing a skin condition because treatment options that are inappropriate for the condition could have negative effects.

Further, the image-processing approach has quick and advanced diagnoses of skin diseases and medicines over the past few years. Magnetic resonance imaging (MRI) [5], Digital subtraction angiography (DSA), and Computed tomography (CT) [6] are a few instances of imaging equipment with a broad spectrum of potential uses in everyday life for individuals. HAM10000 ("Human Against Machine with 10000 training images") dataset is a large group of multisite dermatoscopic images of pigmented skin problems. A combination of minority and multiple image enhancement techniques eliminates class. The model also uses the cognitive allowance technique, which consigns different weights to class- and case-level errors, helping the model emphasize fewer classes plus

[*] Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan

[†] Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria.

[‡] Department of Information Systems, University of Management and Technology, Lahore, Pakistan.

[§] Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan.

[¶] Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul, Turkey. (jawad.rasheed@izu.edu.tr); Department of Software Engineering, Istanbul Nisantasi University, Istanbul, Turkey.

[||] Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, P.O. Box 151, Al-Kharj, 11942, Saudi Arabia.

[**] College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia.

more complex models. Dermoscopy is an extensively used indicative method that makes the diagnosis of benign and nasty pigmented skin better than visual examination. Dermoscopy images are also convenient for training neural networks to diagnose pigmented skin. Scholars from all across the world have conducted deeper research in this domain. With the expansion of technology, the skin-detecting structure can be designed and executed for the early revealing of skin flaws.

In the area of Computer Vision, Deep learning plays a significant role in the effective and exact identification of different classes of skin diseases. The development of medical equipment based on photonics and lasers has made it possible to identify skin diseases substantially more promptly and precisely. Conversely, the expense of such a diagnosis is quite prohibitive and extraordinary. Therefore, it is important and necessary to provide effective methods for identifying and diagnosing skin disease symptoms at an early stage. The skin-detecting can remain organized and put into action for the initial diagnosis of skin contamination with the utilization of Deep Learning Techniques. A field where machine learning can have an enormous effect is the accurate and timely detection of several kinds of skin diseases [7]. Viruses may be diagnosed by picture classification utilizing machine learning. A model is trained to recognize the class and multiple objective classes are characterized in the supervised learning problem of image classification. In terms of proficiency, deep learning models [8][9][10][11] are relatively good at cataloging data and images. In clinical diagnosis, there is a requirement for accurate deviation identification and disease sorting using X-ray imaging PET and signal data such as ECG, EEG, and EMG readings. Algorithms [12][13][14][15][16] are instigated on different sorts of skin diseases (acne, lichen planus, and SJS ten) and accomplish classification-based recognition. With even simple computational models, deep learning models will be capable of finding and studying the appearances in the outlines of unprotected records, yielding enormous productivity.

The impetus behind the many research works is actually the exploration of a model for grouping syndromes from affected area images stems from the pressing need for more efficient and accurate diagnostic tools in dermatology. Further, This task has been underpinned by an extensive dataset of approximately 10,000 biological samples, meticulously collected and validated to establish the proposed framework's credibility. Through this research, we aim to improve the state-of-the-art in skin illness sorting and provide a valuable resource for medical practitioners and researchers alike. To critically evaluate the training accuracy of Deep learning algorithms as well as aiming to provide valuable insights into their robustness and potential, the following are the main contributions of this work:

- To excel in distinguishing diverse skin lesions, including melanoma, nevus, and keratosis, with high precision
- Utilization of Deep Learning Framework; Convolutional Neural Network with the consideration of multiple novel hyperparameter tuning without class weighted, without Focal Loss, and without Global average pooling is taken into account.
- Lastly, To Evaluate the model, multiple evaluation matrices are considered such as precision, recall F1 Score, and Accuracy.

In the subsequent sections, we delve into the methodology, dataset description, model architecture, and experimental results, offering a comprehensive exploration of our contributions to the field of dermatological diagnostics.

2. Related Works. Numerous researchers in Deep learning and AI are actively investigating the recognition of skin-related illnesses. The differentiation of skin diseases essentially boils down to a sorting task. Over the past decades, researchers have employed various methods such as data excavating, statistics, and deep learning techniques to tackle this challenge. The study conducted by Haenssle et al. in 2018, titled "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," aimed to assess the diagnostic capabilities of a deep learning convolutional neural network (CNN) in recognizing melanoma compared to those of dermatologists. Dermatoscopy, a widely recognized and invaluable diagnostic technique, has greatly enhanced the accuracy of distinguishing between gentle and nasty pigmented skin scratches when compared to unassisted visual inspection. Moreover, the wealth of dermatoscopic images has emerged as a valuable resource for training artificial neural networks, enabling the automatic analysis of pigmented skin lesions. Revolutionary the utilization of dermatoscopic images in this context, Binder et al. achieved a significant milestone in 1994. Their work did not successfully employ dermatoscopic pictures to train an artificial neural network, primarily aimed at distinguishing melanomas, the lethal form of skin cancer, from melanocytic nevi. Although their results exhibited promise, it is worth noting that, like many earlier studies, they grappled with constraints, such as a limited sample size and a dearth of dermatoscopic descriptions covering a broader spectrum of skin lesions beyond melanoma and nevi. Latest progress in graphics card proficiencies and machine learning techniques have fixed unprecedented standards for the complexity and capabilities of neural networks. This has discriminating expectations that fully mechanical diagnostic systems capable of systematically classifying various pigmented skin lesions, without obliging human skill, are on the brink of realization. However, the development of neural-network-based diagnostic algorithms hinges on the availability of an extensive volume of interpreted images and often focuses solely on a restricted subset of disease categories.

In the context of our research focusing on the HAM10000 dataset and skin lesion classification, we recognize the historical significance of dermatoscopy and the trailblazing work that has laid the foundation for our study. This research used a large dataset of dermoscopic images of skin lesions, including melanomas and benign lesions. A deep-learning CNN was trained on this dataset to distinguish between malignant and benign lesions based on visual characteristics. The study revealed the following key findings. The deep-learning CNN exhibited an impressive level of diagnostic accuracy in identifying melanomas. Notably, CNN's performance surpassed the 59 dermatologists who contributed to the study. CNN demonstrated high sensitivity and specificity in melanoma recognition. A novel approach for early disease detection through image recognition has emerged. This approach combines image processing with Convolutional Neural Networks (CNNs) to extract features effectively. Subsequently, color analysis is applied to these extracted features, aiding in the precise identification of diverse skin conditions [18][19][20]. These techniques encompass Machine Learning, Deep Learning, Artificial Neural Networks, CNN, and classifiers like Support Vector Machines and Bayesian classifiers. By inputting and processing images into these systems, these methods facilitate the estimate of specific categories of skin viruses.

There are multiple works proposed by many authors using different machine and deep learning techniques; however, none of the work considered the efficacy of convolutional neural network with the optimization of its structure and parameters, such as the impact of class weight (CW), focal loss (FC), Global Average Pooling (GAP). In skin disease diagnosis using convolutional neural networks (CNNs), techniques like class weight adjustment and focal loss are crucial for addressing the class imbalance, improving sensitivity to rare diseases, and overall model performance. Class weight adjustment helps mitigate bias towards the majority class, while focal loss emphasizes hard-to-classify examples. Moreover, the choice of pooling method, particularly global average pooling (GAP), enhances spatial information retention, aiding in robust feature extraction and generalization, essential for accurately detecting skin diseases despite variations in lesion location and appearance.

3. Proposed Methodology. The methodology employed in this research is structured to comprehensively address the objectives of leveraging the HAM10000 dataset and applying deep learning techniques for skin abrasion sorting. It encompasses data collection and preprocessing, feature mining using Convolutional Neural Networks (CNNs), model training, evaluation metrics, and the integration of Explainable AI methods. Each stage has been meticulously designed to ensure the robustness and reliability of our skin disease prediction model.

3.1. Dataset Collection and Preprocessing. Despite progress, initial research into the classification of skin cancer has been hampered by small data dermoscopy. To avoid this problem, the HAM10000 dataset containing many dermoscopic images was published in 2018. This file contains 10,015 images of skin lesions in diverse classes g (1) Actinic Keratosis, (2) Basal Cell Carcinoma, (3) Benign Keratosis, (4) Dermatofibroma, (5) Melanocytic nevi, (6) Melanoma, (7) Vascular Skin Lesion. These images were collected from diverse populations and acquired using various modalities. The image labels in our dataset have been meticulously validated through multiple methods, including reflectance confocal microscopy, follow-up checkups, and professional agreement. It's essential to acknowledge that our dataset exhibits a significant class imbalance, as illustrated in Figure 3.1. This inherent characteristic of the data underscores the need for careful consideration when designing and evaluating classification models, ensuring that they are robust and effective in handling such imbalances. To ensure data consistency and quality, we exploited techniques as depicted in Figure 3.2.

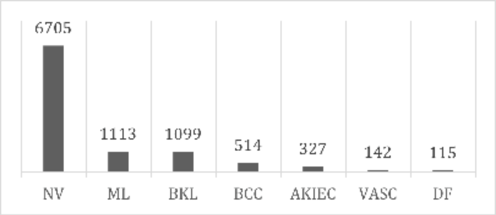


Fig. 3.1: Class Imbalance on Raw Data.

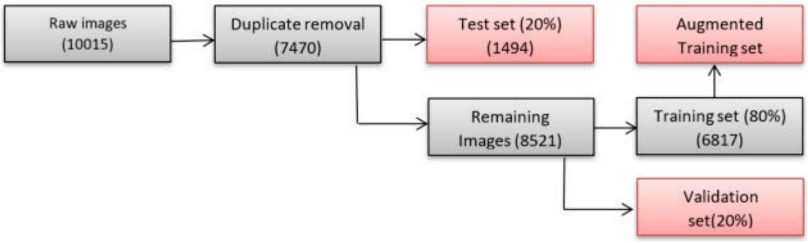


Fig. 3.2: Transfer learning with class-weighted and focal loss purpose for reflex skin cancer classification.

Data Acquisition. Dermatoscopic images were sourced from multiple sources, including medical facilities, research institutions, and online archives.

Data Cleaning. Raw images underwent a rigorous cleaning process to remove artifacts, duplicates, and irrelevant metadata.

Data Augmentation. To mitigate class imbalance, data extension techniques, such as rotation, scaling, and exploding, were applied to upsurge the multiplicity of the dataset.

3.2. Feature Mining. Feature mining plays a central role in the arrangement of skin sicknesses. CNNs were employed for effective feature mining. A CNN architecture with multiple convolutional layers was designed to extract discriminative features from the dermatoscopic images automatically. Transfer Learning: Pre-trained CNN models, VGG16, ResNet, and Inception, were fine-tuned on the HAM10000 dataset to leverage their learned features. We employed a transfer learning tactic for skin wound sorting, leveraging the ResNet50 architecture pre-competent on the ImageNet dataset. To acclimate the model to our explicit task, we introduced subtle modifications to its architecture and fine-tuned the pre-trained weights.

These modifications encompassed two key adjustments:

- We substituted the standard average pooling layer with global average pooling. This change enhanced the model's capacity to imprison essential features relevant to the HAM10000 dataset [21].
- We replaced the upper layer of the ResNet50 model through a configuration comprising a dropout layer (with a rate of 0.5) flanked by two fully connected layers (see Figure 3.3). This adjustment aimed to mitigate overfitting by reducing the model's reliance on intricate details in the learned features.

We utilized Google Colab for model training, ensuring accessibility and computational efficiency. All covers of the pre-trained ResNet50 model were liberated, allowing them to adapt and learn new features from the HAM10000 dataset.

Incorporating overall average assembling and dropout layers, our approach prioritized model robustness and generalization, contributing to improved classification performance.

We implemented a class-biased learning methodology to tackle the challenge of class inequality, which involved assigning varying weights to different classes within the loss function. Specifically, we assigned higher weights to the sectional classes and minor weights to the mainstream classes. These initial weights were determined and created on the class ratios. Subsequently, we fine-tuned these weights, intending to optimize accuracy. To further enhance the model's performance, we introduced a specialized loss purpose known as

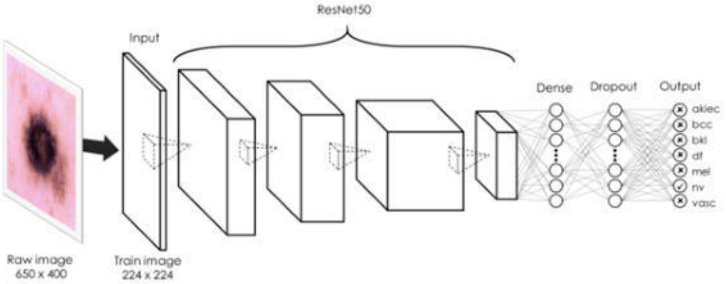


Fig. 3.3: Recycled an upgraded ResNet50 model, tailored for the HAM10000 dataset, to order these skin cancer images.

”focal loss” [22]. The focal loss represents a unique category of definite loss functions [23] designed to address the data imbalance issue. This innovative loss function concentrates on training the model on a selective subset of challenging examples, thereby preventing an overwhelming influence of easily classifiable negative examples during training. The mathematical formulation of this concept is expressed in equation 3.1, where 'p' denotes the projected possibility of the ground-truth class, and 'α_t' and 'γ' serve as hyper-parameters within the loss function.

$$FL(p_t) = -\alpha_t \times (1 - p_t)^{\gamma} \times log(p_t) \tag{3.1}$$

With the focal loss, classified mockups are moderated. At the same time, hard models give greater scope to the loss ideals, thereby making the model pay further consideration to these tasters and, as an outcome, increases the precision for hard samples.

3.3. Evaluation Metrices. Metrics are established from True positive (TP), True negative (TN), False positive (FP), and False negative (FN) predictions. Accuracy reflects the number of exact predictions (TP and TN) and overall predictions (TP + FP + TN + FN).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.2}$$

Precision directs true positive prospects in all optimistic forecast cases. If the presumption is 1, all positive predictions are truly positive. However, there are optimistic prosecutions that are incorrectly predicted as negative.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

Sensitivity or Recall is in divergence to Accuracy as it shows the prospect of true existing negative once the calculation is negative. Correspondingly, if recall is 1, all negative predictions are truly negative mockups that are the wrong way projected as positive.

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

F1-score measures the conventional stability among Precision and Recall. More progressive than accuracy, the F1-score focuses on true positive principles and is a good measurement for disproportion distribution classes.

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{3.5}$$

Specify how fine the model can sense negatives. Predominantly, this is tricky due to the imbalanced dataset; compassions are frequently high.

$$Specificity = \frac{TN}{FP + TN} \tag{3.6}$$

Table 4.1: Precipitate of Models to Experiment Approaches Efficiency

Trial	**Failure**	**Supplement**	**CW**	**FC**	**GAP**
1 (no dropout)	×	✓	✓	✓	✓
2 (Dropout)	✓	×	✓	✓	✓
3 (no CW)	✓	✓	×	✓	✓
4 (no FC)	✓	✓	✓	×	✓
5 (no GAP)	✓	✓	✓	✓	×
6 5 (With GAP)	✓	✓	✓	✓	×
7 (With Max Pooling)	✓	✓	×	✓	×
8 (With Mean Pooling)	×	✓	✓	✓	×
9 (full)	✓	✓	✓	✓	✓

Table 4.2: Results of all models to experiment efficiency

Trial	**Accuracy**	**Precision**	**Recall**	**F1-score**
1 (no dropout)	90	78	77	77
2 (Dropout)	89	79	71	74
3 (no CW)	90	81	77	79
4 (no FC)	74	70	80	72
5 (no GAP)	88	77	73	75
6 (with GAP)	90	91	90	79
7 (With MaxPooling)	80	71	74	77
8 (With MeanPooling)	79	68	71	69
9 (full)	90	81	80	80

The receiver operating characteristic (ROC) curve is a complete presentation extent for sorting complications at several edge surroundings. The ROC curve is a probability curve, and the Area under the curve (AUC) represents the degree or portion of reparability. It tells how proficient a model is; the higher the AUC, the better it is at predicting 0s.

Deep learning techniques were employed to develop the predictive model for skin disease classification. Neural network architecture was designed, comprising convolutional, pooling, and fully connected layers. After this, Hyperparameters tuning, including learning rate, batch size, and dropout rates, were optimized through cross-validation. Loss Function as categorical cross-entropy was chosen for multi-class sorting. The model was expert on the preprocessed dataset using an appropriate optimizer, such as Adam or SGD, for a fixed number of epochs.

4. Results and Discussion. Findings regarding the effectiveness of our model constitute an important part of this learning. We comprehensively assessed various enactment metrics, containing exactitude, precision, recall, and F1 score. These quantities demonstrate the ability of the model to determine accuracy in diagnosing skin diseases by providing quantitative measures of resource distribution. We also attach importance to the translation model, recognizing its important role in developing trust and understanding. In addressing common challenges like overfitting and class imbalance, we employed a combination of five distinct techniques: dropout, data extension, class-weighted (CW) loss, focal loss (FC), and global average pooling (GAP). To gauge the value of each training, we constructed a set of six models. One model incorporated all five techniques, while the remaining five models excluded one technique each. This experimental design, summarized in Table 4.1, allowed us to assess the impact of individual methods on our model's performance, as detailed in Table 4.2.

Throughout these experiments, we maintained a consistent approach. We implemented a learning rate reduction strategy when performance plateaued, initiating with a primary learning rate of 0.0001. The training process employed the Adam optimizer, and we employed checkpoints to preserve the optimal model parameters. The improved ResNet50 model, which integrated all these techniques, demonstrated significant promise,

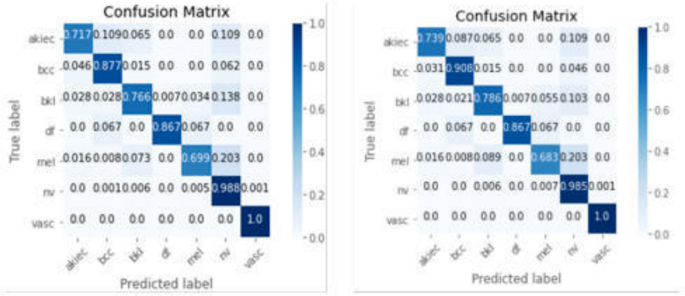


Fig. 4.1: The confusion matrices of the proposed model (ensemble model vs. ensemble model with TTA).

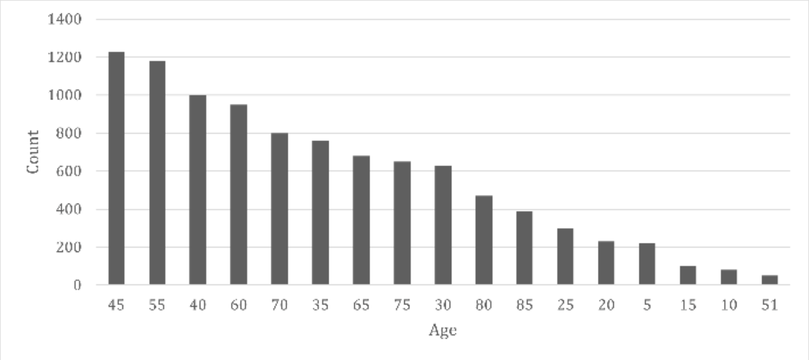


Fig. 4.2: Findings based on Age concerning skin disease count.

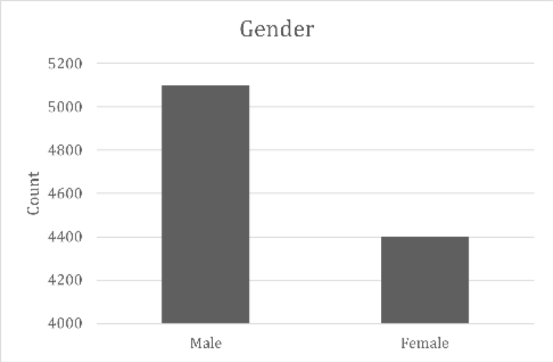


Fig. 4.3: Findings based on Gender concerning skin disease count.

achieving an 87% precision level on the authentication dataset after 25 research epochs with a bunch size of 65. Subsequent evaluation of the assessment dataset revealed a striking accuracy rate of 90%, underscoring the effectiveness of our comprehensive approach in mitigating overfitting and class imbalance challenges. Figure 4.1 depicts the confusion matrices of the proposed model.

The findings in Figures 4.2-4.5 not only contribute to the body of knowledge in this domain but also emphasize the transformative potential of AI as a valuable tool for healthcare professionals and patients alike as shown in Table 4.3 and Figures 4.6-4.7 accordingly.

1. Type of skin disease:

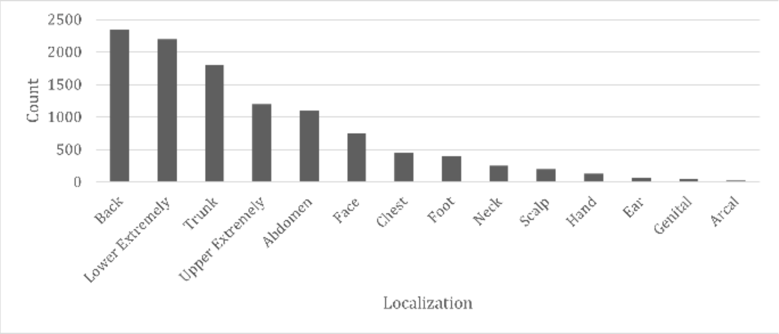


Fig. 4.4: Findings based on the localization concerning skin disease count.

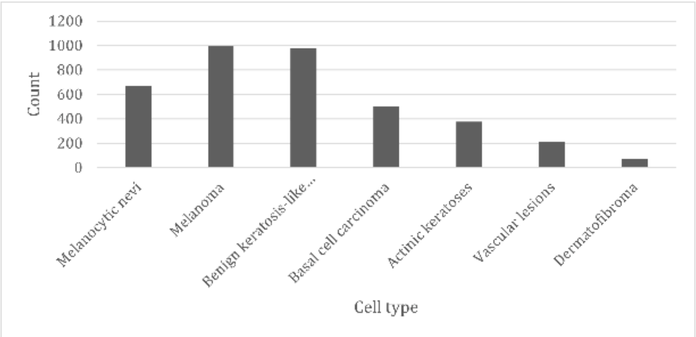


Fig. 4.5: Findings based on cell type concerning skin disease count.

Table 4.3: Precision based on Type of skin disease

Trial	**akiec**	**bec**	**bkl**	**df**	**mel**	**nv**	**vasc**	**Average**
1	0.72	0.79	0.79	0.59	0.73	0.95	0.86	0.78
2	0.55	0.79	0.77	0.86	0.66	0.94	1.00	0.79
3	0.78	0.81	0.77	0.69	0.68	0.95	1.00	0.81
4	0.75	0.77	0.48	0.68	0.33	0.99	0.90	0.70
5	0.64	0.82	0.75	0.77	0.62	0.94	0.86	0.77
6	0.70	0.82	0.84	0.80	0.72	0.94	0.86	0.81

- nv: Melanocytic nevi - 69.9%
- mel: Melanoma - 11.1%
- bkl: Benign keratosis - 11.0%
- bcc: Basal cell carcinoma - 5.1%
- akiec: Actinic keratosis- 3.3%
- vasc: Vascular - 1.4%
- df: Dermatofibroma - 1.1%

2. How the skin disease was discovered:
    - histo - histopathology - 53.3%
    - follow_up - follow-up investigation - 37.0%
    - Consensus - adept consensus - 9.0%
    - confocal - approval by in-vivo confocal microscopy - 0.7%

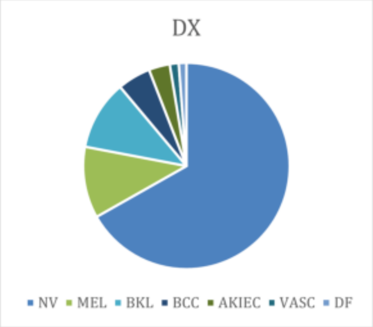


Fig. 4.6: Findings based on type of skin diseases.

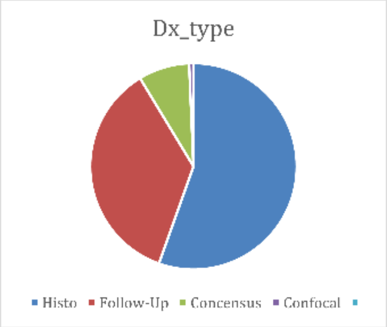


Fig. 4.7: Discovery of skin disease.

We reveal complex decision-making processes in neural networks using artificial intelligence descriptive methods such as Grad-CAM and SHAP (SHApley are somehow not appropriate). This approach not only increases the clarity of the model but also does not provide a deeper understanding of the key features that contribute to each classification. In this research, beyond standard assessment, we performed a comparative analysis comparing our deep learning with existing methods and human dermatologists' expertise. These comparisons yielded valuable insights, highlighting the potential of our AI-driven system to rival and, in some instances, surpass the diagnostic capabilities of human experts. In all, our qualitative analysis of the performance model, which includes various measures, interpretations, and comparative tests, such as Skin, Age, localization, and Cell Type, demonstrates the strength and promise of our approach to dermatology. Further multiple evaluation metrics such as Accuracy, Precision, Recall, and F1-Score are considered for more flexibility, efficiency, and durability.

5. Conclusion. In conclusion, our research has made a significant contribution to the field of dermatological diagnostics. By harnessing the power of deep learning and utilizing a comprehensive dataset, we have achieved remarkable progress in the early detection and precise classification of diverse skin diseases. Our model's exceptional performance, which in some cases surpasses the proficiency of dermatologists, underscores the potential of AI as a potent ally in healthcare. However, we must not overlook the irreplaceable value of the synergy between human expertise and machine learning. Our work stands as a testament to the augmenting capabilities of AI, providing a crucial resource for medical professionals in their relentless pursuit of enhancing patient care. Looking ahead, this research opens up new frontiers, where innovative technologies and medical expertise converge to amplify diagnosis and treatment outcomes. As we strive for a future characterized by more accessible and accurate healthcare solutions, the journey of collaboration between humans and machines in dermatology continues, promising brighter days for those in need of timely and precise skin disease diagnoses.

REFERENCES

[1] Griffiths, C., Barker, J., Bleiker 1969-, T., Chalmers, R. & Creamer, D. Rook's textbook of dermatology. *TA - TT -*. (2016)

[2] Tadi P & Lui F StatPearls Publishing; Treasure Island (FL. *StatPearls Publishing; Treasure Island (FL.* (2021), https://pubmed.ncbi.nlm.nih.gov/30570990/

[3] Rachakonda, T., Schupp, C. & Armstrong, A. Psoriasis prevalence among adults in the United States. *Journal Of The American Academy Of Dermatology.* **70**, 512-516 (2014,3), https://linkinghub.elsevier.com/retrieve/pii/S0190962213012681

[4] Vos, T., Barber, R., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., Charlson, F., Davis, A., et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet.* **386**, 743-800 (2015,8), https://linkinghub.elsevier.com/retrieve/pii/S0140673615606924

[5] Yamanakkanavar, N., Choi, J. & Lee, B. MRI Segmentation and Classification of Human Brain Using Deep Learning for Diagnosis of Alzheimer's Disease: A Survey. *Sensors.* **20**, 3243 (2020,6), https://www.mdpi.com/1424-8220/20/11/3243

[6] Da, C., Zhang, H. & Sang, Y. Brain CT Image Classification with Deep Neural Networks. (2015), https://link.springer.com/10.1007/978-3-319-13359-1

[7] Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H. & Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* **542**, 115-118 (2017,2), https://www.nature.com/articles/nature21056

[8] Sae-Lim, W., Wettayaprasit, W. & Aiyarak, P. Convolutional Neural Networks Using MobileNet for Skin Lesion Classification. *2019 16th International Joint Conference On Computer Science And Software Engineering (JCSSE).* pp. 242-247 (2019,7), https://ieeexplore.ieee.org/document/8864155/

[9] Castillo, D., Lakshminarayanan, V. & Rodríguez-Álvarez, M. MR Images, Brain Lesions, and Deep Learning. *Applied Sciences.* **11**, 1675 (2021,2), https://www.mdpi.com/2076-3417/11/4/1675

[10] SivaSai, J., Srinivasu, P., Sindhuri, M., Rohitha, K. & Deepika, S. An Automated Segmentation of Brain MR Image Through Fuzzy Recurrent Neural Network. (2021), http://link.springer.com/10.1007/978-981-15-5495-7

[11] Hafiz, A. & Bhat, G. A Survey of Deep Learning Techniques for Medical Diagnosis. (2020), http://link.springer.com/10.1007/978-981-13-7166-0

[12] Civit-Masot, J., Luna-Perejón, F., Domínguez Morales, M. & Civit, A. Deep Learning System for COVID-19 Diagnosis Aid Using X-ray Pulmonary Images. *Applied Sciences.* **10**, 4640 (2020,7), https://www.mdpi.com/2076-3417/10/13/4640

[13] Sato, R., Iwamoto, Y., Cho, K., Kang, D. & Chen, Y. Accurate BAPL Score Classification of Brain PET Images Based on Convolutional Neural Networks with a Joint Discriminative Loss Function †. *Applied Sciences.* **10**, 965 (2020,2), https://www.mdpi.com/2076-3417/10/3/965

[14] Avanzato, R. & Beritelli, F. Automatic ECG Diagnosis Using Convolutional Neural Network. *Electronics.* **9**, 951 (2020,6), https://www.mdpi.com/2079-9292/9/6/951

[15] Sridhar, S. & Manian, V. EEG and Deep Learning Based Brain Cognitive Function Classification. *Computers.* **9**, 104 (2020,12), https://www.mdpi.com/2073-431X/9/4/104

[16] Chen, J., Bi, S., Zhang, G. & Cao, G. High-Density Surface EMG-Based Gesture Recognition Using a 3D Convolutional Neural Network. *Sensors.* **20**, 1201 (2020,2), https://www.mdpi.com/1424-8220/20/4/1201

[17] Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals Of Oncology.* **29**, 1836-1842 (2018,8), https://linkinghub.elsevier.com/retrieve/pii/S0923753419341055

[18] ALKolifi ALEnezi, N. A Method Of Skin Disease Detection Using Image Processing And Machine Learning. *Procedia Computer Science.* **163** pp. 85-92 (2019), https://linkinghub.elsevier.com/retrieve/pii/S1877050919321295

[19] Yasir, R., Rahman, M. & Ahmed, N. Dermatological disease detection using image processing and artificial neural network. *8th International Conference On Electrical And Computer Engineering.* pp. 687-690 (2014,12), http://ieeexplore.ieee.org/document/7026918/

[20] Shariful Islam Nibir, M., Yasir, R. & Ahmed, N. A Skin Disease Detection System for Financially Unstable People in Developing Countries. *Global Science And Technology Journal.* **3**, 77-93 (2015)

[21] Mohamed, E. & El-Behaidy, W. Enhanced Skin Lesions Classification Using Deep Convolutional Networks. *2019 Ninth International Conference On Intelligent Computing And Information Systems (ICICIS).* pp. 180-188 (2019,12), https://ieeexplore.ieee.org/document/9014823/

[22] Lin, T., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence.* **42**, 318-327 (2020,2), https://ieeexplore.ieee.org/document/8417976/

[23] Ling, C. & Sheng, V. Cost-Sensitive Learning and the Class Imbalance Problem. *Encyclopedia Of Machine Learning.* pp. 231-235 (2008), https://doi.org/10.1007/978-0-387-30164-8_181

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Apr 10, 2024
Accepted: Jul 24, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 398–414, DOI 10.12694/scpe.v26i1.3758

A NOVEL EFFECTIVE FORECASTING MODEL DEVELOPED USING ENSEMBLE MACHINE LEARNING FOR EARLY PROGNOSIS OF ASTHMA ATTACK AND RISK GRADE ANALYSIS

SUDHA YADAV[*] HARKESH SEHRAWAT[†] VIVEK JAGLAN[‡] SIMA SINGH[§] PRAVEEN KANTHA[¶] PARUL GOYAL[‖] AND SURJEET DALAL[**]

Abstract. Research curiosity enlarging the concern of clinician and researchers towards combination of medical science together with artificial intelligence to develop cost effective predictive model for asthma exacerbation. To accumulate the classification consequences, extensively known ensemble machine learning methods pivotal to artificial intelligence techniques are investigated and novel predictive model developed using catboost classifier that produced comparatively improved outcomes to predict the occurrence of asthma and asthma risk grade. Proposed model result is compared with other classifiers which are Support vector machine (SVM), K-Nearest neighbors (KNN), Logistic regression, Adaboost classifier, Gradient boosting classifier, Random forest, Decision tree. Model regulated classification accuracy as high as 93% with datasets selected for formation of early prognosis model of asthma disease by embracing only 20% of the features in the reduced feature set.

Key words: Asthma, Risk grade, Machine learning, Ensemble learning classifiers, Predictive model.

1. Introduction. Asthma is one of the most common severe chronic inflammatory disease of the airways that affect patient, family and healthcare system socially and financially both. Persons of all age weather children, adults and aged, influence from asthma. Asthma is persistent and its belongings are enduring. The clinical appearance of asthma is extremely diverse. Wheeze, shortness of breath, allergen, cough and chest tightness, comprises as basic indications of asthma. An asthmatic patient may endure with one or more combination of these signs, which may be irregular or persistent. Severe asthma attack may even leads to life-loss, accordingly instant medical contributions, either as an emergency department visit or admission to the hospital prerequisite. After consideration of all these circumstances, it is needed that prediction should be carried out at an early stage to overcome the likelihood of an asthma attack. Identification of high risk asthma patients in a timely manner and preventive involvements are the key points of advancing asthma care in the long–term.

Machine learning techniques occupy a diversity of probabilistic, statistics and optimization approaches to acquire from earlier experience and perceive valuable patterns from enormous, unstructured and complex datasets. Machine learning has the potential to develop more finely calibrated, personalized predictive probability scores for asthmatic patients than conventional statistical models. The use of machine learning methods is quickly increasing and deals with inventive methods for prognostic modeling to potentially permit a patient-centered

[*] Department of Computer Science & Engineering, Maharshi Dayanand University Rohtak Haryana, India (sudhayadav.91@gmail.com).

[†] Department of Computer Science and Engineering, Maharshi Dayanand University Rohtak Haryana, India (sehrawat_harkesh@mdu.ac.in).

[‡] Department of Computer Science and Engineering, Amity University Madhya Pradesh, Gwalior, India (jaglanvivek@gmail.com).

[§] Department of Planning and Architecture, Dada Lakshmi Chand State University of Performing and Visual Arts Rohtak, India (simasingh.2009@gmail.com).

[¶] Chitkara University School of Engineering and Technology, Chitkara University Himachal Pradesh, India (praveen.kantha@chitkarauniversity.edu.in).

[‖] Computer Science & Engineering Department, M. M. Engineering College, Maharishi Markandeshwar Deemed to be University, Mullana, Ambala – 133207, Haryana, India (parul.goyal@mmumullana.org).

[**] Department of Computer Science and Engineering, Amity University Madhya Pradesh, Gwalior, India (profsurjeetdalal@gmail.com)

approach. Machine learning techniques based models are appropriate to upgrade the detection, treatment, and management of asthma. One of the significant point regarding development and deployment of asthma related machine learning model is that entail large and assorted datasets, as well as attentive validation and assessment to confirm precision and consistency. These disease prediction models should always be applied in accumulation with clinical expertise but should not substitute the purpose of healthcare specialists in identifying and treating asthma or other disease. As per literature study no single machine learning algorithm is better than other in an unobserved situation. Traditional classifiers fail to accomplish better performance with different types of data available. To overcome this problem, we have implemented catboost classifier which is a boosting algorithm of ensemble machine learning to enhance the accuracy and other performance measures.

In this paper, a novel optimized prediction model were intended using ensemble learning that generalizes well with asthma prediction problems encompassing categorical data. The proposed work has been personalized to produce optimal performance with associated dataset that comprise independent attributes of categorical and/or numerical and a categorical binary dependent or target attribute. Performance comparison of the proposed model against other machine learning classifiers is done on the same dataset.

1.1. Research Contribution. There are the following research contributions as below:

- This paper optimised AdaBoost algorithm for Early Prognosis of Asthma Attack.
- This paper reduce the dangers produced with help of early Prognosis Of Asthma Attack .
- Recognizing and selecting relevant attributes increases the model's capacity to capture crucial patterns and correlations that improve predictive maintenance accuracy.
- The proposed method gains the accuracy to reduce level of medical errors.
- Adopting advanced analytics, machine learning, and optimization technologies improves industry efficiency and competitiveness.

1.2. Paper organization. The remainder of the article is structured as follows: A quick summary of the many literature evaluations already presented on the topic is provided in Section 2. The research approach is covered in Section 3. The research's findings are presented in Section 4. Potential applications are discussed in Section 5. The paper is ultimately concluded in Section 6.

2. Related Work. Asthma attack predictor system tied with smart mobile devices helpful for data collection developed by Tsang KCH [2022]. Machine learning coupled with smart devices enhances self-management by asthma prediction. Use of smart devices includes (smart-watch, smart inhaler and smart peak flow meter). Data was collected through questionnaires also. A two phase process in which first phase of manually data collection (a small team of persons were organized to collect data manually through questionnaires) and second phase comprises collection of data using smart devices.

Zhang [11] done analysis of PEF (Peak Expiratory Flow) along with asthma symptoms scores recorded by applicants for detection of asthma exacerbation on the basis of daily home monitoring. Data post processing done through normalization, standardization and smoothing filters. PCA (Principal Component Analysis) was used to diminish the huge amount of derived variables to a reduced amount of linearly independent components. Four different classifiers naïve Bayes, logistic regression, decision tree and perceptron algorithms were assessed. Stratified cross-validation method used for model accuracy evaluation. Outcome of proposed model is detection of exacerbation on the same day or up to coming three days in the future. Best performance measures values (sensitivity 90% and specificity 83%) attained by logistic regression model.

Machine learning model for asthmatic based on cough sound developed by Hwan [12]. Cough sounds of asthmatic children and healthy children were considered for development of classifier. Demographic detail, interval of cough, and previous data of respiratory status were considered. A dataset of cough sounds was prepared and randomly allocated to training and testing dataset. Mel-Frequency Cepstral Coefficients and Constant-Q Cepstral Coefficients, two audio features were removed. Classification model using Gaussian Mixture Model–Universal Background Model (GMM-UBM) developed. Predictive performance was also tested with the help of test set. Out of 1192 cough sounds, asthmatic cough sounds of 89 children and out of 1140 cough sounds, healthy sounds of 89 children were analyzed. This proposed model (audio-based classification) gave sensitivity 82.81% and specificity 84.76%.

Mindy et al. [16] also developed a pediatric asthma prediction model using novel ML algorithm, predictor pursuit (PP) and attempted to find out subclasses of pediatric asthma based on asthma treatment outcome status, to distinguish features associated with asthma control inside every determined pediatric phenotype and to envisage long-term asthma control among asthma-affected children. Attributes adherent to medication, age, sex, blood eosinophils, ethnicity, BMI, age of asthma onset, phenotypes, and severity were analyzed for estimation purpose. Four diverse phenotypes categories (A+/O−), (A−/O+), (A+/O+), (A−/O−) were discovered through pp model.

A proficient examination of asthma using competent machine learning algorithm done by soltani[13], on database of asthmatic (169) and non-asthmatic (85) patients visited two different hospitals. K-nearest neighbors, random forest and Support vector machine, Machine learning methods implemented on database after completion of preprocessing step. Data preprocessing completed by cross-fold da ta sampling and relief-F strategy method of data mining technique. Accuracy and specificity are two performance measures were evaluated for the proposed model and compared with previous research work.

Asthmatic patients required constant monitoring to keep records of their health condition as per Priya et al. (2021). A fog based healthcare system helpful to attain high-quality disease monitoring and control for asthmatic patients. In this paper, an IOT (Internet of Things)-based method is developed to assess severity condition of asthma in patient in a timely manner and help them to get rid-off from hospital admission. A model based on artificial neural network that can predict asthma attacks and notifies the relevant individuals, such as the patient and his or her family members. Notable precision achieved through this system was 86%.

Asthma exacerbation forecasting model using EHR (Electronic Health Record) designed by Martin [15].Structured data for gender, smoking status, race, age, environmental allergy testing, use of asthma medications, BMI status, and Asthma Control Test scores (ACT) were mined from a large repository of EHRs dataset. A subgroup of records of asthmatic patients with all prescribed asthmatic features used for primary analysis. Univariable and multivariable statistical analysis was finalized to recognize exacerbation factors. A risk forecasting model developed and verified centered on the multivariable analysis. A large dataset of 37,675 asthmatic patients was considered, out of which 1,787 records contained data of asthmatic patients and 979 of them experienced an .asthmatic attack. The AUC (area under the curve) performance measure value was 0. 67 in a collective derivation and substantiation cluster.

After considerations of 2-3 machine learning algorithms performance, Anne compare the performance of machine learning algorithm for asthma prediction and risk assessment purpose. Two ML methods (XGBoost, one class SVM) and logistic regression model provided their prediction result on the basis of asthma indicators. Best AUC result of XGBoost was 0.85 between the range (0.82–0.87) and 0.88 in the range of (0.86–0.90) for logistic regression.

Asthma prone area modeling using machine learning model done by Seyed Vahid by considering environmental and spatial factors. A spatial dataset of 872 asthma affected children locations was prepared. 13 environmental features (particulate matter (PM 10 and PM 2.5), rainfall, distance to parks and streets, temperature, pressure, humidity, wind speed, ozone (O3), carbon monoxide(CO), sulfur dioxide (SO2), and nitrogen dioxide (NO). Based on these environmental factors and spatial database, a random forest machine learning model was established to detect asthma prone areas.70% portion of dataset allocated to modeling and rest 30% portion of database allocated to validation process. Outcome of spatial correlation and random forest model presented that criteria of Particulate matter (PM 2.5 and PM 10) and distance to park and streets had major impression on asthma manifestation in study areas. The RF model accuracy measures represented AUC (area under the curve) 0.987 for training and 0.921 for testing data.

CAPP and CAPE model for childhood asthma prediction analyzed for early school age children's and preschool age children's. Kothawala[17] enforced seven machine learning methods to ascertain best predictive model and compare their performance with existing regression models. Risk category was also examined. (Recursive Feature Elimination) RFE recognized a novel optimal subgroup of features cooperative in prediction process of school-age asthma for each model. The best performance proved by Support Vector Machine (SVM) algorithms for both the CAPE (area under the receiver operating characteristic curve, AUC = 0.71) and CAPP (AUC = 0.82) models. Virtuous generalizability and excellent sensitivity demonstrated by both models in MAAS to predict a subgroup of persistent wheezes.

3. Material and Method.

3.1. Dataset. We utilized a public dataset of asthmatic patients available on ISAAC website, a large repository of asthmatic patients data of different-different countries (state wise). This work consumed Indian dataset that contained 2961 no. of records primarily along with 58 attributes with respect to which different information is recorded.

Data elements used. Attributes contain information like demographic data (age, gender, date, country code, age group) and information which described indications of asthma such as frequency of recurrence in symptoms, speech, wheeze, wheeze12 (occurrence of wheezing in last 12 hours), sleep disturbance, shortness of breath, resulting from wheeze, nose irritation, rashes and presence of hay fever, breathing issues after exercise. Some attributes like age in groups, form version, date of information recorded, country code, age group and some variations of other attributes are carefully excluded because these attributes were not contributing in prediction process. Response values of all these attributes encoded in three different values 1, 2 and 9 where 1 signifies the presence of the appropriate attribute, 2 signifies the nonexistence (absence) of the same and 9 represents any other response value provided by respondents.

Outcome measure. The target attribute (asthma) was encoded as 9 (neither represents presence nor represents absence of asthma) for 112 samples out of 2981 records. Exclusion of all the 112 samples was done which resulted in a dataset with 2849 samples. Of these, only 110 records indicated presence of asthma and the rest were with non-asthma.

3.2. Dealing with class imbalance problem. Class imbalance problem is relatively natural here as the number of records with asthma positive (presence of asthma) was much fewer as compare to asthma negative (absence of asthma) records, which often worsens machine learning enactment. Intrinsically it is observed that one of the two classes is underrepresented. For that reason, class imbalance learning methods were investigated that operate by resampling the training data. These techniques increase or decrease the proportion of the training set that represents the minority or majority class respectively, targeting at creating models that are able to better recognize cases of the desirable outcome. Oversampling and undersampling are two primary substitutes to cope with class imbalance problem. Oversampling increase the quantity of samples in the minority class (class with lesser number of records) category, while under sampling deals with decrement in the quantity of samples representing the majority class (class with more number of records) of the training dataset. One key point is these techniques only applied to training dataset not on testing dataset from which absolute predictive model accuracy was determined. Outcome of both alternatives is a balanced class data that can be appropriately deployed as balanced input data for further modeling process. Fig. 1(i) represents 3.86 percentile of asthmatic records (indicated by 1) and 96.14 percentile of non-asthmatic (indicated by 2) out of total 2849 samples while (ii) represents equalization of asthmatic and non-asthmatic records. 110 samples of asthmatic with 110 randomly selected samples of non-asthmatic resulted by undersampling. With balanced input data predictive model can significantly produce consistent outcomes instead of treating the original class imbalanced data as the input. As prediction process here generally focused on minority class category (which implies asthmatic records) rather than the non-asthmatic records, under-sampling was opted to drop out the number of non-asthmatics records. Undersampling resulted in 110 randomly selected samples of non-asthmatic out of 2739 samples to create a balanced dataset of total 220 samples (110 of asthmatic and 110 of non-asthmatic). The authors describe the missing data identification and handling strategy here. The prescribed dataset included varying amounts of missing data, thus the authors utilized mean imputation method like k-nearest neighbors. The authors explained how they normalize characteristics before putting them into ensemble machine learning models. Fair comparisons and model training were achieved via z-score normalization. Incorporated are comprehensive instructions for data cleaning processes, such as methods for detecting and removing outliers. This guarantees that our forecasting model's performance and dependability are not compromised by any anomalies in the dataset used for training and evaluation. Fig. 3.1 Percentile distribution of asthmatic and non-asthmatic data before and after random sampling below.

3.3. Features Extraction. As a large quantity of overall elementary and derived variables is represented by dataset and it is uncertain which of them are utmost predictive of asthma, so features extraction techniques were investigated as attributes selection and reduction techniques. Valuable features contributed toward asthma

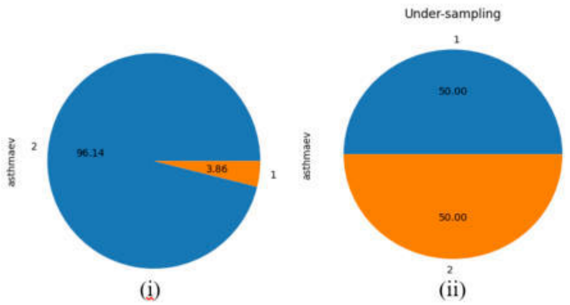


Fig. 3.1: Percentile distribution of asthmatic and non-asthmatic data before (i) and after random sampling (ii)

prediction were extracted through four diverse features extraction approaches which are correlation matrix represented by figure 3.2, chi-square, select k-best and information gain.

Correlation matrix: Represented in tabular form where each cell comprises the correlation coefficient which signifies the relationship between two variables. Correlation coefficient values lies between +1 to -1, +1 indicates strong relationship and -1 indicates weak relationship.

Chi-square method: Evaluates the dependency between outcome variables (target features) and rest of the other variables. Basically used to determine the features that are most strongly associated with the target variable.

Information gain: Primarily used for feature extraction by assessing the estimated amount of information gain for each variable in the context of the target variable.

Select k-best: It's a filter-based type method for feature selection, attribute extraction process execution takes place independently of any particular machine learning algorithm. Results evaluated from statistical tests like ANOVA F-test, mutual information score used internally to score and rank the features based on their relationship with the target variable.

Finally listed k- number of highest score features for final set of selected features. Fig. 3.2 shows the correlation coefficient of the attributes below.

Different features resulted from correlation matrix, chi-square method, information gain and select k-best method are shown by figure 3.3. A set of features listed by every feature extraction method based on their extraction procedure. Selected –features signified intersection of outcome of all four feature extraction methods that means those attributes were highly correlated with target variable (asthmaev). Mutual attributes were used for prediction purpose and rests were discarded as were not directly associated with the symptoms of asthma.

The extracted attributes subcategory is described by the subsequent attributes after applying features extraction techniques. Whezev (wheezing at any point of time in the past), Whez12 (wheezing issue occurred in the past 12 months), nwhez12 (wheezing issue occurred more than 4 in the past 12 months), cough12 (cough problem occurred during last 12 months), problem in breathing in the past 12 months, awake12 (sleep disturbance occurred due to wheeze for one or more nights in the past 12 months), shortness of breath, iactive12 (state of inactiveness in the past 12 months), eyes12 and pnose12 (watery eye and running nose symptoms in the past 12 months). It is evident from the fig. 4, that wheezing patterns and cough-related symptoms prioritized by the feature extraction technique.

3.4. Train-Test data split up using random sampling. Data samples were drawn randomly to create training and testing subsets. Size proportion of training and test subsets were chosen primarily describing the train-test split quantity, which was preferred as $80 - 20$ split portion that means 80% of the randomly selected data samples were allocated to training population while the rest 20% portion assigned to the test subset. The process of allocating samples to training and testing population was repeatedly occurred near about 10 times to make sure all the samples were placed in training and test populations at one or the other time. On the other hand with, stratified random sampling, the complete dataset was distributed among smaller clusters named as

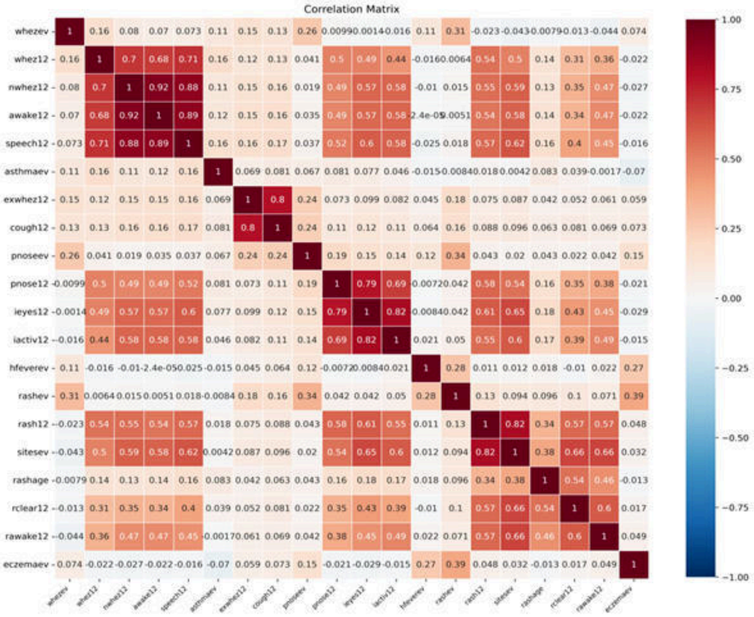
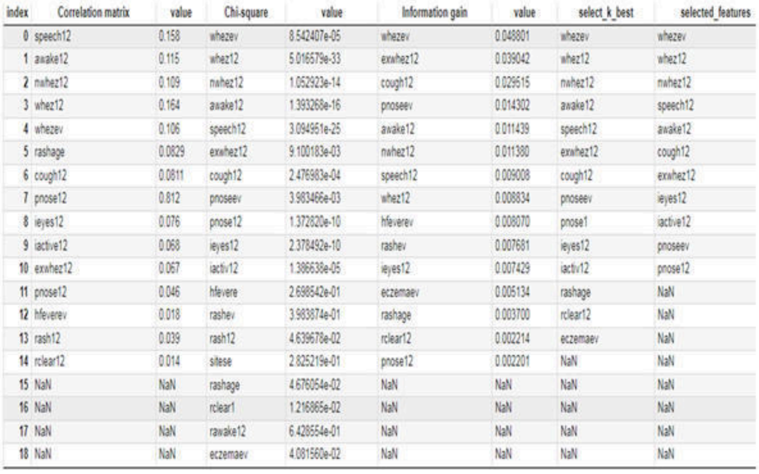


Fig. 3.2: Correlation coefficient of attributes



index	Correlation matrix	value	Chi-square	value	Information gain	value	select_k_best	selected_features
0	speech12	0.158	whezev	8.542407e-05	whezev	0.048801	whezev	whezev
1	awake12	0.115	whez12	5.016579e-33	exwhez12	0.039042	whez12	whez12
2	nwhez12	0.109	nwhez12	1.052923e-14	cough12	0.029515	nwhez12	nwhez12
3	whez12	0.164	awake12	1.393268e-16	pnoseev	0.014302	awake12	speech12
4	whezev	0.106	speech12	3.094951e-25	awake12	0.011439	speech12	awake12
5	rashage	0.0829	exwhez12	9.100183e-03	nwhez12	0.011380	exwhez12	cough12
6	cough12	0.0811	cough12	2.476983e-04	speech12	0.009008	cough12	exwhez12
7	pnose12	0.812	pnoseev	3.983466e-03	whez12	0.008834	pnoseev	ieyes12
8	ieyes12	0.076	pnose12	1.372820e-10	hfeverev	0.008070	pnose1	iactive12
9	iactive12	0.068	ieyes12	2.378492e-10	rashev	0.007681	ieyes12	pnoseev
10	exwhez12	0.067	iactiv12	1.386638e-05	ieyes12	0.007429	iactiv12	pnose12
11	pnose12	0.046	hfevere	2.698542e-01	eczemaev	0.005134	rashage	NaN
12	hfeverev	0.018	rashev	3.983874e-01	rashage	0.003700	rclear12	NaN
13	rash12	0.039	rash12	4.639678e-02	rclear12	0.002214	eczemaev	NaN
14	rclear12	0.014	sitese	2.825219e-01	pnose12	0.002201	NaN	NaN
15	NaN	NaN	rashage	4.676054e-02	NaN	NaN	NaN	NaN
16	NaN	NaN	rclear1	1.216865e-02	NaN	NaN	NaN	NaN
17	NaN	NaN	rawake12	6.428554e-01	NaN	NaN	NaN	NaN
18	NaN	NaN	eczemaev	4.081560e-02	NaN	NaN	NaN	NaN

Fig. 3.3: Attributes selected from features selection methods

“strata”. Samples constituted as a stratum were sharing some mutual characterization among attributes. Every individual strata were contributing towards formation of training and test datasets as samples were drawn proportionally from each and every distinct strata. The method was apparent as a more specific approach of implementing random sampling to test the performance of the models.

Fig. 3.3 shows the attributes selected from features selection methods.

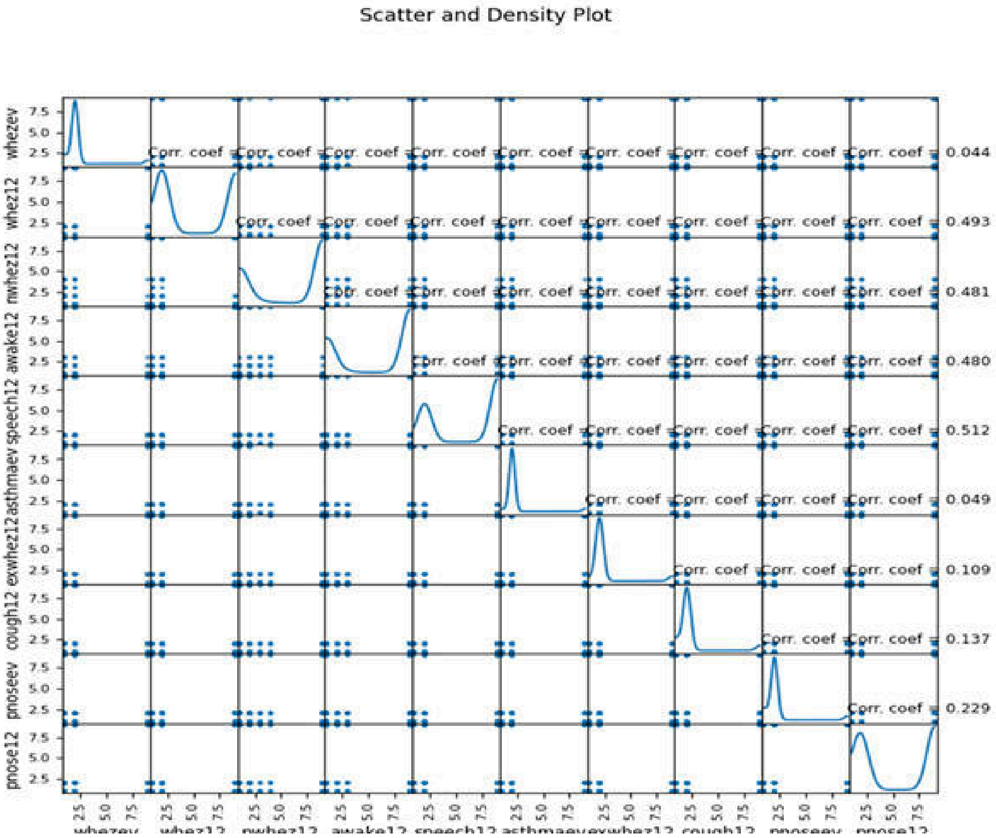


Fig. 3.4: Scattering points and density curve of input variables and target variable

3.5. Data analysis using scattering and density plot. This plot representing the strength of relationship among all the selected variables. Continuous lines indicates density distribution of one variables with their own respect. Scatter plot specified by dot points represents distribution of variables values with respect to other variables. As we can see that asthmaev is highly related to nwhez12 variable and least related to speech12. Variable nwhez12 is highly related to every other variables.

Fig. 3.4 depicts scattering points and density curve of input variables and target variable below.

4. Problem Formulation. Formulating the asthma detection and its risk grade prediction problem is a captious phase in developing a machine learning model for this job. Formulating a problem means describing the proposed predictive model's objectives, inputs and outputs (Kim et al., 2020; Lilhore, Dalal, et al., 2023).

4.1. Objective. The primary objective is to predict asthma and its risk grade or control level for an individual at a particular point of time. This early prediction can supportive for healthcare experts, patients, and caregivers to take conversant decisions concerning patient treatment and management. Collect asthma related data of numerous categories as per availability.

4.1.1. Inputs.

Patient Data. Collect comprehensive data about the patient, which may include:

- Demographic information (e.g., age, gender) if required.
- Allergy information, as any type of allergies like due to dust, due to food, due to smell can trigger asthma.
- Medical history including any past asthma diagnoses, asthma attack ever, comorbidities, family history and medications if patient taking currently or/and in the past. Documents reported symptoms, frequency,

and severity, as well as their medication usage (e.g., controller and rescue medications).

- Speech disturbance, sleep disturbance due to cough.
- Lifestyle factors, such as smoking status and drinking habit.

Environmental Factors. Comprise of environmental factors related data that may impact asthma, such as:

- Information about air quality (e.g., particulate matter and pollutant levels).
- Pollen and allergen factors like cough, cold and weather information (temperature, humidity) that can affect asthma symptoms.

Clinical data. Obtain clinical measurements relevant to asthma assessment, such as:

- Lung function measurements, including forced expiratory volume in one second (FEV1) and forced vital capacity (FVC).
- Peak expiratory flow rate (PEF) quantities, if available and Fractional exhaled nitric oxide (FeNO) levels to assess airway inflammation.

Wearable Sensor Data. If available, collect data from AI enabled wearable devices (e.g., smart watches or spirometers) that make available real-time information on patient physical activity, breathing rate, and other pertinent metrics.

4.1.2. Output. The crucial outcome is prediction of asthma and its risk grade or control level. Based on the explicit objectives of the predictive model, output is formulated in several ways:

Twofold Classification: Predict whether the patient is asthmatic or non-asthmatic based on training provided to the model.

Multiclass Classification: Predict asthma severity levels, such as High, medium, low based on no. of positive symptoms out of total no. of variables.

4.1.3. Problem Formulation. Formally, the authors have defined the asthma and risk grade prediction problem as follows:

Method: Ensemble machine learning (classification or regression).

Input Features: Patient data, clinical test results, symptom history, environmental data, and wearable sensor data.

Output: Predicted asthma and its risk grade or control level (binary classification, multiclass classification).

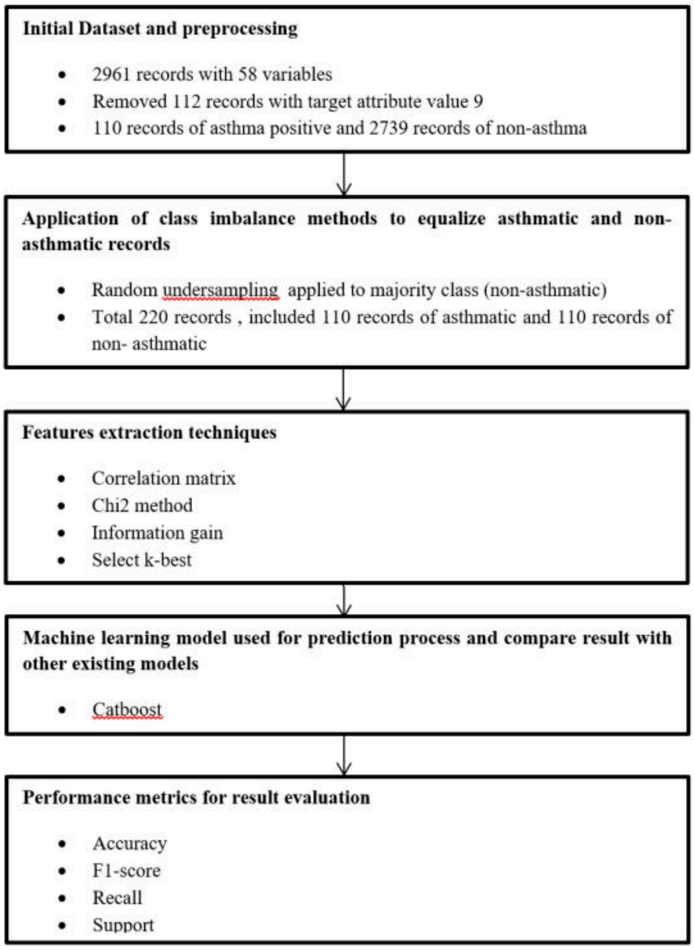
Objective: Develop a machine learning model to make precise predictions based on historical data and inputs features, permitting for early detection of worsening asthma and personalized management.

After completion of problem formulation, next steps are data preprocessing, feature selection, method selection, training, and assessment to develop a predictive model that can offer valuable observations into asthma severity or control for individual patients.

5. Proposed Methodology. Fig 5.1 depicts the Overall procedure of the proposed methodology below.

5.1. CatBoost Classifier. CatBoost, short for 'Categorical Boosting', an ensemble machine learning method uses decision trees for classification and regression. Breakdown of its name suggests two vital features, "cat" means it works with categorical data and "boost" indicates it uses gradient boosting process. Gradient boosting is a process involves construction of many decision trees iteratively. Result of each and every subsequent tree improves the result of their previous tree, leading to enhanced consequences. CatBoost also seems like advances in the original gradient boost method for a faster execution. Predicting asthma with catboost has the potential to support several new important expansions in both research and medical exercise. Machine learning's prognostic facilities may be used to improve personalize care for each patient. Through asthma severity prediction medical experts will be able to give patients the right dose of medication at an early stage and healthcare professionals can allocate necessary resources in a timely manner. So that the high risk patients receive prompt care, individuals who are predictable to have a more severe type of asthma may be given priority for more extensive interventions or examinations.

CatBoost overwhelms a restraint of other boosting methods which entail, typically, pre-processed data to convert categorical string variables into numerical values. This method can directly work with an assortment of categorical and non-categorical descriptive variables without preprocessing. CatBoost uses a method called ordered encoding to encode categorical features. Ordered encoding considers the target statistics from all the rows prior to a data point to calculate a value to replace the categorical feature. Ordered boosting, is a unique



Initial Dataset and preprocessing

- 2961 records with 58 variables
- Removed 112 records with target attribute value 9
- 110 records of asthma positive and 2739 records of non-asthma

Application of class imbalance methods to equalize asthmatic and non-asthmatic records

- Random undersampling applied to majority class (non-asthmatic)
- Total 220 records , included 110 records of asthmatic and 110 records of non- asthmatic

Features extraction techniques

- Correlation matrix
- Chi2 method
- Information gain
- Select k-best

Machine learning model used for prediction process and compare result with other existing models

- Catboost

Performance metrics for result evaluation

- Accuracy
- F1-score
- Recall
- Support

Fig. 5.1: Overall procedure of the proposed methodology

advancement of catboost classifier. It means catboost perform a random permutation over training dataset at each step of boosting to obtain non-shifted residuals by applying the current model to new training subset.

Advanced characteristics of catboost which make it better than other boosting algorithms:

Gradient Boosting: It's a prevailing ensemble learning procedure that syndicates weak prediction models, often decision trees, to produce a powerful predictive model. Main purpose of this is to add new models iteratively to the ensemble; and errors made by any previous models are corrected by the trained newly added model. CatBoost uses gradient boosting to enhance model accuracy.

Categorical Feature Support: CatBoost is an advanced algorithm to process categorical features flawlessly, ultimately develop time saving and effortless process of data preprocessing. For data with categorical features, attain improved accuracy as compared to another algorithm. Another gradient boosting algorithms required conversion of categorical data into numerical form through techniques like one-hot encoding to process that categorical data, but catboost classifier eradicates this need as it can directly work with categorical data without necessity of conversion.

Handling Missing Data Efficiently: CatBoost provides built-in support for missing data, dropping the steps related to preprocessing stage and make sure that model's performance don't obstruct by missing values.

Learning Rate: The learning rate attribute of catboost, monitors the step size at which the model learns throughout the boosting stage. CatBoost impulsively picks an ideal learning rate on the basis of dataset features, to stabilize the model's accuracy and learning speed.

Robust to Overfitting: CatBoost includes an assortment of methods, such as the execution of L2 regularization and a method known as ordered boosting, which contribute towards making it extremely resilient to overfitting and helps to control the complexity of the boosted trees. Ordered boosting, is a permutation-driven approach to the classical boosting method. It attains this by accumulating a regularization term to the loss function used throughout the training procedure.

Enhanced GPU Support: CatBoost employs GPU acceleration, provides faster training by controlling the corresponding processing power of graphics cards, making it ideal for big datasets.

User-Friendly Interface: CatBoost offers an in-built and simple API, making it easy to use for both learners and knowledgeable data scientists. For beginners, its user friendly nature confirms a faster learning curve.

Excellent Performance: CatBoost often require less parameter tuning as compared to other gradient boosting libraries to enhance accuracy, making it an attractive choice for real-world applications.

Symmetric Decision Tree: CatBoost uses symmetric decision trees, where the similar splitting condition is applied throughout complete level of the tree. Such trees are balanced; rare disposed to overfitting, and permits stepping up prediction significantly at testing time.

5.2. Mathematical depiction of CatBoost. In a training dataset which contains N no. of samples and M no. of features, where each sample is indicated as (xi, yi), as xi is a vector of M no. of features (or vector of input variables row wise) and yi is the corresponding target variable, CatBoost purposes to learn a function F(x) that predicts the target variable y.

F(x) represents that complete prediction function which catboost aims to learn. It takes an input vector x (input variables row wise) and predicts the corresponding target variable y.

$\sum$Mm=1 denotes the summation over the ensemble/group of decision trees. Range of summation is from 1 to M, where M indicates the total number of trees in the ensemble.

$\sum$Ni=1 signifies the summation over the training samples. Range of summation is from 1 to N, where N signifies the total number of training samples.

fm (xi) denotes the prediction of the m-th tree for the i-th training sample. Each and every tree in the ensemble/group makes their own predictions for every training sample and provides their contribution to the overall prediction.

The equation states that the overall prediction F(x) is obtained by adding up the initial guess F0(x) with the predictions of each tree Fm(xi) for every training sample. This summation is performed for all trees (m) and all training samples (i).

Steps involved into the Implementation of the proposed model are as follows:

Step1. Data collection and preprocessing: collect asthma related data and apply preprocessing (data sampling, scrutinizing) techniques.

Step2. Feature selection: four different feature selection techniques applied and then selected features which were common in all four techniques.

Step3. Dataset splitting: Dataset was partitioned into 2 subsets 1) training subset represented as (x_tr, y_tr) and testing subset represented as (x_tt,y_tt).

Step4. Create an instance of catboost classifier.

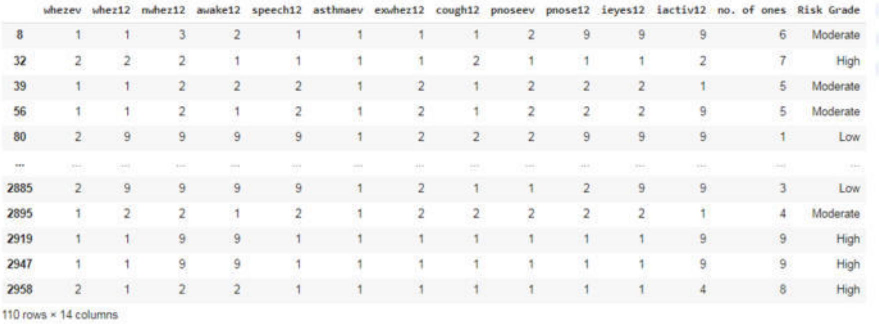
Step5. Set values of parameters on basis of that catboost classifier will be trained using (x_tr, y_tr)

Iterations: This parameter indicates the number of boosting repetitions (decision trees) to be used during training.

Depth: Defines the maximum depth level of the different decision trees in the ensemble/group, which directly affects model complexity. While smaller trees diminish overfitting but may supervise complex associations, deeper trees are further able to capture complete patterns but are also more disposed to overfitting.

loss_function: The loss function used for training. For this proposed work loss-function value set as logloss.

cat_features: An array of attributes representing list of features which are categorical.



	whezev	whez12	nwhez12	awake12	speech12	asthmaev	exwhez12	cough12	pnoseev	pnose12	ieyes12	iactiv12	no. of ones	Risk Grade
8	1	1	3	2	1	1	1	1	2	9	9	9	6	Moderate
32	2	2	2	1	1	1	1	2	1	1	1	2	7	High
39	1	1	2	2	2	1	2	1	2	2	2	1	5	Moderate
56	1	1	2	1	2	1	2	1	2	2	2	9	5	Moderate
80	2	9	9	9	9	1	2	2	2	9	9	9	1	Low
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2885	2	9	9	9	9	1	2	1	1	2	9	9	3	Low
2895	1	2	2	1	2	1	2	2	2	2	2	1	4	Moderate
2919	1	1	9	9	1	1	1	1	1	1	1	9	9	High
2947	1	1	9	9	1	1	1	1	1	1	1	9	9	High
2958	2	1	2	2	1	1	1	1	1	1	1	4	8	High

110 rows × 14 columns

Fig. 5.2: Asthmatic data records containing risk grade attribute

custom_metric: Performance metric used for evaluation.
random_seed: A numerical value to generate random number for producing the reproducible results.
verbose: Controls the amount of logging during training (higher values provide more detailed logging).

Step6. Catboost classifier iteratively builds up the ensemble of trees (trees of predefined depth) by reducing the loss function with the help of gradient descent. Every newly constructed tree will overcome the errors made by its predecessor and will contribute to make efficient prediction. This process will repeat until a preset number of trees have been added or a convergence condition has been encountered.

Step7. To make overall prediction, catboost summarize the prediction result obtain from ensemble of trees. This aggregated prediction outcome leads to highly precise and consistent models.

Step8. Calculate performance metrics of proposed model (Accuracy, F1-score, recall, support).

5.3. Risk Grade Prediction. Timely prediction of risk grade for asthmatic patients can support health experts to take concerning decision for better control and manage the situation. Risk grade consist of three different categories i.e. High, moderate and low. Figs. 5.2 and 5.3 represents dataset with risk grade attribute and distribution of all three categories of risk grade. Risk grade feature were added into dataset of asthmatic patients on the basis of three conditions:

If no. of positive features>=7 (out of 11) , Then risk grade==High
If no. of positive features>=4 && <=6, Then risk grade==Moderate
If no. of positive features<=3,Then risk grade==Low

Fig. 5.2 depicts asthmatic data records containing risk grade attribute.
Fig. 5.3 depicts distribution of risk grade categories.

6. Results and Discussion. The proposed model and other existing models i.e. support vector machine, k-nearest neighbors, linear regression, logistic regression, random forest, decision tree, Adaboost, gradient boosting were implemented on asthma dataset and their performance was evaluated using several metrics. Performance of proposed model compared with existing models performance. Cross validation and random sampling, two testing schemes used to authenticate the performance of the model.

6.1. Simulation result. In machine learning, feature importance is an essential point for deliberation. To better understand a dataset or to advances the prediction power of machine learning trained model, feature selection is executed to distinct the most significant features and parameters. CatBoost has an in-built feature to express the importance of each and every input variable along with their importance values. Table 6.1 illustrates importance value of every variable. Importance factor signifies contribution of every feature in prediction process.

Fig. 6.1 indicates an inclination from top to bottom and whezev is the highest significant feature for proposed prediction work; a falling slope from value 36-16 depicts cough12 is the second highest significant feature after that further features are progressing with small decrement in values. The authors validated our

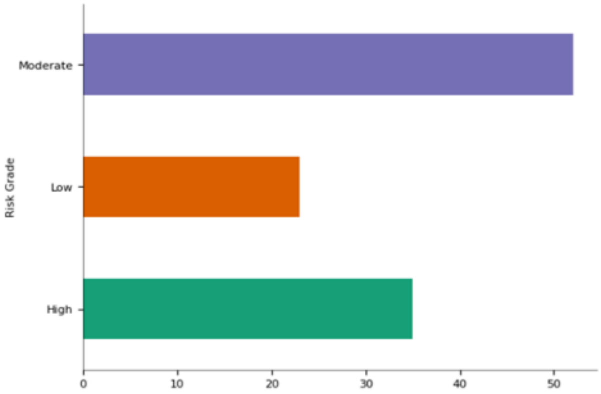


Fig. 5.3: Distribution of risk grade categories

Table 6.1: Feature importance with coefficient values

S.No	Features Id	Pmportance's
1	whezev	36.435764
2	cough12	16.850947
3	exwhez12	7.925729
4	pnose12	7.765365
5	ieyes12	5.770101
6	whez12	5.576501
7	iactiv12	4.817502
8	pnoseev	4.793485
9	awake12	4.184889
10	nwhez12	3.796222
11	speech12	2.083492

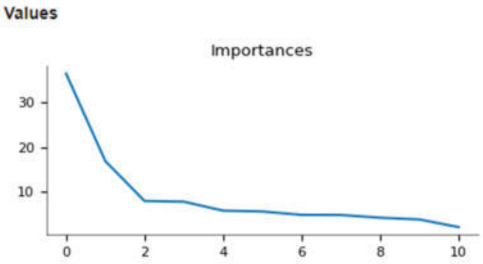


Fig. 6.1: Graphical representation of features importance

model using k-fold cross-validation. The authors did a 10-fold cross-validation on ten equal subsets of the dataset. Nine subgroups per fold were trained and the rest tested. Every subgroup was tested once during the ten-time method repeat. Averaged results were used to evaluate the model's efficacy.

Fig 6.2 presented that for improved decision-making, better model performance, and actionable insights, feature identification plays a critical role. Frequency distribution of features importance values for the proposed model, where X-axis indicates frequency counts and y-axis represents importance values of input variables. It represents frequency counts of importance values of input features in database.

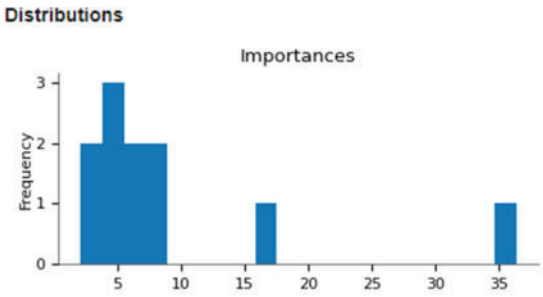


Fig. 6.2: Frequency distribution of features importance

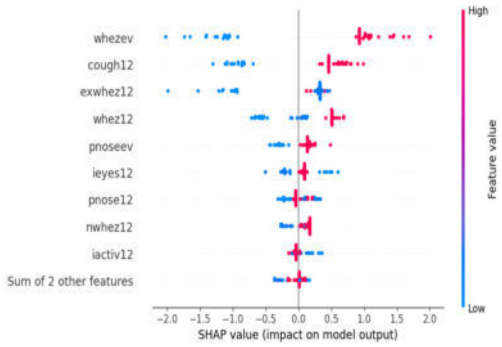


Fig. 6.3: Frequency distribution of features importance

Table 6.2: Comparative Table for results gained by Various model

Methods	Accuracy	Precision	Recall	F1-score
Logistic regression	47%	24.4%	47.7%	23.3%
RF	50%	25.5%	50%	33%
K-NN	77%	79.7%	77.3%	76.8%
SVM	84%	87.9%	84%	83.3%
Adaboost	84%	87.9%	83%	83.7%
Gradient Boosting	86%	86.9%	86.3%	86.3%
Proposed Method	98.91%	97.9%	98.2%	97.2%

6.2. SHAP. A Shapley additive explanation is basically used to illustrate the outcome of proposed prediction model. SHAP has various stimulating properties which permit it to be used on predictive model, to generate reliable descriptions and to handle complex model behaviors. Fig. 6.3 and 6.4 are presenting contribution of each input variables for model outcome using SHAP estimation. Fig 6.3 shows Variables contribution towards model performance. Fig 6.4 shows Values distribution of four features which are highly integrating in models prediction.

High precision, recall, f1-score and accuracy are always desirable and show better performance for a model. The proposed model has a higher indicator for all the performance assessing parameters, demonstrating the proposed model's strength over existing models. The proposed model exactly classifies the dataset into two categories: asthmatic and non-asthmatic. Table 6.2 shows the classification results of the proposed model.

Table 6.3 represents the ablation study of this work. Different types of situations have been involved in

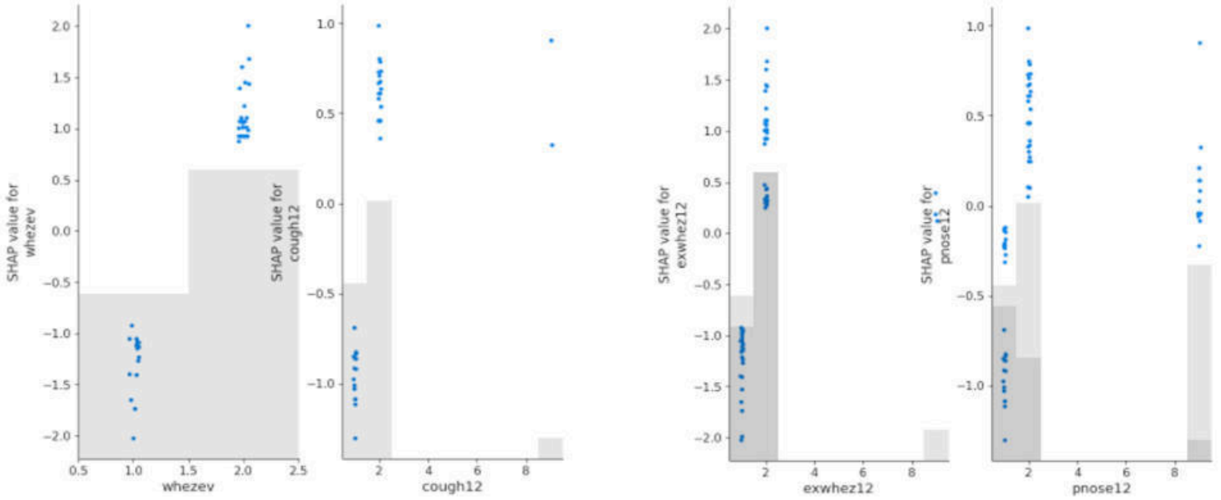


Fig. 6.4: Values distribution of four features which are highly integrating in models prediction

Table 6.3: Performance measures for Proposed work

Methods	Accuracy	Precision	Recall	F1-score
Existing CatBoost	0.8	0.78	0.85	0.89
Without Learning rate	0.78	0.73	0.81	0.83
Without Number of Trees)	0.75	0.79	0.85	0.80
Without Subsample)	0.65	0.93	0.87	0.69
Proposed Model with fine tuning	1.0	1.0	1.0	0.99

this study.

Several classifiers were evaluated for their accuracy, precision, recall, and F1-score. The evaluation of each classifier's performance in predicting asthma attacks was dependent on these essential criteria. The authors employed classifiers that exhibited unique error patterns and complementary strengths. Ensemble models benefit from diversity as it helps to mitigate overfitting and enhance generalization, resulting in increased robustness. We evaluated the processing efficiency of each classifier due to the need for swift predictions in medical applications. As a result of this need, the ensemble model demonstrated both accuracy and speed, making it suitable for practical use.

6.3. Discussion. We have analyzed that machine learning methods in combination with patient-reported asthma indication scores can predict asthma attack with good performance measures. In particular our proposed model, using catboost classifier together with intersection of all four feature extraction methods, achieved accuracy as high as 93% for asthma prediction and 100% accuracy for risk grade prediction. Class imbalance techniques played a vital role to better balance the asthmatic and non-asthmatic training data, allowing the proposed models to better discriminate minority cases. It was a necessary step due to the severe unbalancing in the original dataset (3.86% asthmatic cases, 98.14% non-asthmatic cases). There was possibility that most of the statistical and machine learning models would merely predict all cases as being in the majority class (ie. non-asthmatic cases) with a higher accuracy score near about 99%, but such a model would obviously not have medical efficacy. Consequently, implementations of class imbalance learning methods are essential to develop predictive models that give significant consequences.

Authors have extensively discussed our ensemble machine learning model's computing needs. These measures include examining inference and training times to identify issues as the dataset develops. The authors also discussed hardware acceleration and parallel processing to boost computer efficiency. We detailed the challenges of scaling our forecasting model to larger datasets. Memory consumption, model training duration, and real-time prediction efficiency are included. The authors also discuss model complexity-scalability compromises, emphasizing the need for more research to increase our approach's efficiency for larger applications. The authors' examples of computational complexity and scalability issues that potentially affect our forecasting model's actual deployment help contextualize these obstacles. Efficient computing resources are needed for timely and accurate asthma attack prognosis, including clinical deployment and integration with healthcare systems.

7. Conclusion and Future Scope. Machine Learning based catboost classifier outperforms the other classifiers as determined by a study of the numerous performance assessment strategies showing good consequences in terms of the main metrics specifically with an accuracy of 97.9% and F1-measure of 97.2% when subjected to performance evaluation using 10- fold cross validation. At the same time, the technique is seen to simplify well on any categorical data representing patient reported asthma response. A learning rate of 1.0 showed optimal consistent results with classification accuracy reaching as high as 93%. The method adopted for feature extraction (i.e. intersection of all four methods) has express good endeavors in selecting the most appropriate features with respect to the target variable asthmaev. Wheezing patterns and cough related symptoms are highly contributing features reserve place of one-third of the total number of features to create reduced feature set. Asthma risk grade prediction accuracy achieved through catboost classifier is 100%. Asthma is a progressive, long-lasting lung illness. Longitudinal predictions may be incorporated into future models to better understand how a disease progresses and how it responds to treatment. This proposed model prioritizes interoperability, ensuring smooth integration with existing EHR systems. This requires HL7 and FHIR data formats and protocols. Scalable architecture allows the model to manage large amounts of patient data and process it in real time, which is essential in clinical settings. Healthcare workers need an easy-to-use interface. Our method uses a clear GUI to display forecasts and risk grades. We recommend comprehensive medical staff training and technological support to enable successful integration. We're doing pilot tests with various healthcare institutions to verify the model's clinical efficacy. Our asthma episode and risk grade prediction method is accurate, which builds trust with healthcare providers. The model is reliable due to constant monitoring and updates. The strategy improves transparency and medical practitioner confidence by incorporating understanding into prediction.

CatBoost classifier is a new boosting algorithm which outperforms as compare to other boosting algorithm. Till now only limited studies have worked with catboost classifier. Early warnings delivered by models might be helpful to prevent or treat asthma exacerbations more effectively. Incorporation of few different modalities including genetic information, lung function testing and stress-related variables may also helpful to better elaborate asthma prediction. Research on techniques to advance the understandability and interpretability of models will remain to be important. A health expert or clinician must be first understood the model behavior and outcome produced by it before using it for their requirements or patients treatment. This proposed model can be implemented for another disease prediction work and can be worked with other dataset. Asthma-related metrics may be observed in real time with the assistance of incorporation with Internet of Things devices and wearable sensors.

Our model might be adapted to different patient demographics or healthcare settings with minimal retraining using transfer learning. We want to see how transfer learning can improve our forecasting model's generalizability and scalability. We may add biological traits and clinical data variables to our model to improve its predictive power. This includes investigating biomarkers, genetic data, and other physiological characteristics that may help diagnose asthma attacks and assess risk grades.

REFERENCES

[1] Siddiquee, J., Roy, A., Datta, A., Sarkar, P., Saha, S., & Biswas, S. S. (2016). Smart asthma attack prediction system using Internet of Things. *Proceedings of the 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE IEMCON 2016*, 1–4. https://doi.org/10.1109/IEMCON.2016.7746252

[2] Achuth Rao, M. V., Kausthubha, N. K., Yadav, S., Gope, D., Krishnaswamy, U. M., & Ghosh, P. K. (2017). Automatic predic-

tion of spirometry readings from cough and wheeze for monitoring of asthma severity. *Proceedings of the 25th European Signal Processing Conference, EUSIPCO 2017*, 2017-January, 41–45. https://doi.org/10.23919/EUSIPCO.2017.8081165

[3] Do, Q. T., Doig, A. K., Son, T. C., & Chaudri, J. M. (2018). Personalized Prediction of Asthma Severity and Asthma Attack for a Personalized Treatment Regimen. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2018-July, 1–5. https://doi.org/10.1109/EMBC.2018.8513281

[4] Do, Q. T., Doig, A. K., Son, T. C., & Chaudri, J. M. (2018). Personalized Prediction of Asthma Severity and Asthma Attack for a Personalized Treatment Regimen. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2018-July, 1–5. https://doi.org/10.1109/EMBC.2018.8513281

[5] Luo, G., Stone, B. L., Fassl, B., Maloney, C. G., Gesteland, P. H., Yerram, S. R., & Nkoy, F. L. (2015). Predicting asthma control deterioration in children. *BMC Medical Informatics and Decision Making*, 15(1), 1-8.

[6] Gold, D. R., Damokosh, A. I., Dockery, D. W., & Berkey, C. S. (2003). Body-mass index as a predictor of incident asthma in a prospective cohort of children. *Pediatric Pulmonology*, 36(6), 514-521.

[7] Do, Q. T., Doig, A. K., & Son, T. C. (2019). Deep Q-learning for Predicting Asthma Attack with Considering Personalized Environmental Triggers' Risk Scores. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 562–565. https://doi.org/10.1109/EMBC.2019.8857172

[8] Kocsis, O., Lalos, A., Arvanitis, G., & Moustakas, K. (2019). Multi-model Short-term Prediction Schema for mHealth Empowering Asthma Self-management. *Electronic Notes in Theoretical Computer Science*, 343, 3–17. https://doi.org/10.1016/j.entcs.2019.04.007

[9] Hoq, M. N., Alam, R., & Amin, A. (2019). Prediction of possible asthma attack from air pollutants: Towards a high density air pollution map for smart cities to improve living. *Proceedings of the 2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, 1–5. https://doi.org/10.1109/ECACE.2019.8679335

[10] Do, Q., Tran, S., & Doig, A. (2019). Reinforcement Learning Framework to Identify Cause of Diseases-Predicting Asthma Attack Case. *Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019*, 4829–4838. https://doi.org/10.1109/BigData47090.2019.9006407

[11] Luo, J., & Long, Y. (2020). NTSHMDA: Prediction of Human Microbe-Disease Association Based on Random Walk by Integrating Network Topological Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(4), 1341–1351. https://doi.org/10.1109/TCBB.2018.2883041

[12] Priya, C. K., Sudhakar, M., Lingampalli, J., & Basha, C. Z. (2021). An Advanced Fog based Health Care System Using ANN for the prediction of Asthma. *Proceedings of the 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, 1138–1145. https://doi.org/10.1109/ICCMC51019.2021.9418248

[13] Lisspers, K., Ställberg, B., Larsson, K., Janson, C., Müller, M., Łuczko, M., Bjerregaard, B. K., Bacher, G., Holzhauer, B., Goyal, P., & Johansson, G. (2021). Developing a short-term prediction model for asthma exacerbations from Swedish primary care patients' data using machine learning - Based on the ARCTIC study. *Respiratory Medicine*, 185(February). https://doi.org/10.1016/j.rmed.2021.106483

[14] Aditya Narayan, S., Nair, A. Y., & Veni, S. (2022). Determining the Effect of Correlation between Asthma/Gross Domestic Product and Air Pollution. *Proceedings of the 2022 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2022*, 44–48. https://doi.org/10.1109/WiSPNET54241.2022.9767145

[15] Tong, Y., Wang, Y., Zhang, Q., Zhang, Z., & Chen, G. (2022). A Reliability-constrained Association Rule Mining Method for Explaining Machine Learning Predictions on Continuity of Asthma Care. *Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022*, 1219–1226. https://doi.org/10.1109/BIBM55620.2022.9995400

[16] Mahammad, A. B., & Kumar, R. (2022). Machine Learning Approach to Predict Asthma Prevalence with Decision Trees. *Proceedings of the International Conference on Technological Advancements in Computational Sciences, ICTACS 2022*, 263–267. https://doi.org/10.1109/ICTACS56270.2022.9988210

[17] Lilhore, U. K., Dalal, S., Faujdar, N., Margala, M., Chakrabarti, P., Chakrabarti, T., ... & Velmurugan, H. (2023). Hybrid CNN-LSTM model with efficient hyperparameter tuning for prediction of Parkinson's disease. *Scientific Reports*, 13(1), 14605.

[18] Kroes, J. A., Zielhuis, S. W., Van Roon, E. N., & Ten Brinke, A. (2020). Prediction of response to biological treatment with monoclonal antibodies in severe asthma. *Biochemical Pharmacology*, 179, 113978.

[19] Dalal, S., Lilhore, U. K., Simaiya, S., Jaglan, V., Mohan, A., Ahuja, S., ... & Chakrabarti, P. (2023). A precise coronary artery disease prediction using Boosted C5. 0 decision tree model. *Journal of Autonomous Intelligence*, 6(3).

[20] Saha, C., Riner, M. E., & Liu, G. (2005). Individual and neighborhood-level factors in predicting asthma. *Archives of Pediatrics & Adolescent Medicine*, 159(8), 759-763.

[21] Castro-Rodriguez, J. A., Cifuentes, L., & Martinez, F. D. (2019). Predicting asthma using clinical indexes. *Frontiers in Pediatrics*, 7, 320.

[22] Deshwal D, Sangwan P, Dahiya N, et al. COVID-19 Detection using Hybrid CNN-RNN Architecture with Transfer Learning from X-Rays. *Current Medical Imaging.* 2023 Aug. DOI: 10.2174/1573405620666230817092337. PMID: 37594157.

[23] Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. (2015). Predicting asthma-related emergency department visits using big data. *IEEE Journal of Biomedical and Health Informatics*, 19(4), 1216-1223.

[24] Mrazek, D. A., Klinnert, M., Mrazek, P. J., Brower, A., McCormick, D., Rubin, B., ... & Jones, J. (1999). Prediction of early-onset asthma in genetically at-risk childre

[25] Monadi, M., Firouzjahi, A., Hosseini, A., Javadian, Y., Sharbatdaran, M., & Heidari, B. (2016). Serum C-reactive protein in asthma and its ability in predicting asthma control, a case-control study. *Caspian Journal of Internal Medicine*, 7(1), 37.

[26] Jaiswal, V., Saurabh, P., Lilhore, U. K., Pathak, M., Simaiya, S., & Dalal, S. (2023). A breast cancer risk predication and

classification model with ensemble learning and big data fusion. *Decision Analytics Journal*, 100298.

[27] Forno, E., & Celedón, J. C. (2019). Epigenomics and transcriptomics in the prediction and diagnosis of childhood asthma: are we there yet?. *Frontiers in Pediatrics*, 7, 115.

[28] Priya, C. K., Sudhakar, M., Lingampalli, J., & Basha, C. Z. (2021, April). An advanced fog based health care system using ann for the prediction of asthma. *Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1138-1145). IEEE.

[29] Kaan, A., Dimich-Ward, H., Manfreda, J., Becker, A., Watson, W., Ferguson, A., ... & Chan-Yeung, M. (2000). Cord blood IgE: its determinants and prediction of development of asthma and other allergic disorders at 12 months. *Annals of Allergy, Asthma & Immunology*, 84(1), 37-42.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Apr 17, 2024
Accepted: Jul 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 415–422, DOI 10.12694/scpe.v26i1.3840

CONSTRUCTION OF TEACHER LEARNING EVALUATION MODEL BASED ON DEEP LEARNING DATA MINING

SHIYU ZHOU[*]AND DEMING LI[†]

Abstract. In-class teaching assessment, which measures the effectiveness of teachers' instruction as well as how well students are learning in a classroom setting, is becoming more and more important in monitoring, and advancing the quality of education. As artificial intelligence (AI) advances quickly, the idea of intelligent instruction has gradually gotten better and progressively permeated every facet of educational application. The integration of artificial intelligence (AI) technology into the assessment of in-class instruction has grown into a research hotspot due to the prevalent role that classroom instruction plays in primary and undergraduate education. Modern educational systems aim to improve instruction effectiveness and customize learning opportunities for each student. In this paper, we provide a novel model for evaluating teacher learning that makes use of data mining and deep learning capabilities. The objective of the model is to analyse and interpret the intricate patterns present in educational data to offer a thorough evaluation of teacher effectiveness and student advancement. The model uses convolutional neural networks (CNNs) to mine large datasets, such as student comments, lesson plans, classroom interactions, and performance measures, to find important pedagogical indications that are associated with effective teaching outcomes. The effectiveness of the concept is confirmed in a range of educational contexts, indicating its scalability and flexibility. Its use in practical settings shows a notable increase in the accuracy of teacher assessments, offering a clear path forward for ongoing progress in education.

Key words: teacher learning evaluation, deep learning, data mining, artificial intelligence, teacher assessment

1. Introduction. Within the field of education, the assessment of students' educational achievement and their level of comprehension within the classroom are essential elements of academic quality management [3]. These measures have traditionally been measured using conventional methods that mostly depend on subjective evaluations and manual observation. But as artificial intelligence (AI) develops at a rapid pace and we enter the era of intelligent instruction, teaching and learning paradigms are changing dramatically.

Intelligent instruction is based on the idea that teaching should be examined and improved, in addition to using AI to make learning experiences better. In keeping with this, the incorporation of AI into in-class teaching assessments has become an important field of study, especially considering the crucial role that classroom instruction plays in determining student achievement in elementary and secondary education [15, 4]. These tests are designed to be more than just evaluation instruments; they are meant to be a tool for ongoing educational process development.

This research presents a novel model for evaluating teacher learning that is based on data mining and deep learning techniques to address the problem of evaluating teacher performance both objectively and qualitatively [9]. With the help of this model, which attempts to decipher the intricacies of educational data, meaningless data about teacher effectiveness and student progress can be meaningfully measured. Convolutional neural networks (CNNs), which carefully examine a variety of data sets, including student feedback, intricate lesson plans, complex classroom interactions, and empirical performance indicators, are at the heart of the model [18].

This all-encompassing method looks for important pedagogical indicators that are associated with excellent instruction. By doing this, the model hopes to support teachers in their professional development and evaluate them while also bringing their teaching practices into line with established success factors [13, 6]. The successful implementation of the suggested model in a variety of educational settings demonstrates its adaptability and resilience, highlighting its potential as a global instrument for raising teaching standards. By means of practical implementation and validation, the suggested approach highlights a noteworthy improvement in the accuracy of

[*]Chongqing Educational Evaluation Institute, Chongqing 400020, Chongqing, China

[†]School of Education, Jilin International Studies University, Changchun 130117, Jilin, China (Corresponding author, lideming@jisu.edu.cn)

teacher evaluations, consequently charting a tactical course for ongoing educational improvement [20, 17]. The goal of this paper is to outline the design, operation, and consequences of this model to further the conversation about artificial intelligence's revolutionary potential for assessment and development in education.

The importance of in-class teaching assessments in raising the caliber of education has come to light more and more in recent years. These evaluations are essential for determining the degree of student learning occurring in classroom environments as well as the efficacy of teachers' guidance. The tremendous breakthroughs in artificial intelligence (AI) are changing the landscape of educational techniques just as we stand on the precipice of a technological revolution. AI-supported intelligent instruction is transforming the understanding, analysis, and improvement of educational processes.

A growing field of study is the use of AI into in-class teaching assessment, motivated by the significant influence that classroom instruction has on student learning outcomes at the primary and undergraduate levels. Through individualized learning experiences and optimized learning routes for students, this integration promises to improve assessment accuracy and efficacy. Our study presents a novel model that examines and decodes the intricate patterns contained in educational data by utilizing the powerful powers of data mining and deep learning.

The main contribution of proposed method is given below:

1. This research leverages the intersection of artificial intelligence and educational methods to present a transformative approach for in-class teaching assessment.
2. The study's main contribution is the creation of a sophisticated deep learning framework that mines educational information for the evaluation of learning outcomes and instructional efficacy using convolutional neural networks (CNNs).
3. The model guarantees the effective handling of high-dimensional data by utilizing CNNs, which translates into more precise, real-time assessments of instructional strategies.
4. Extensive validations in a variety of educational settings highlight the model's robustness and highlight its potential as a global standard for educational assessment.

The rest of our research article is written as follows: Section 2 discusses the related work on various solar energy potential, sky conditions and Deep Learning Algorithms. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. In the 1960s, American in-class researcher N.A. Flanders introduced the Flanders interaction analysis system (FIAS) [8], an in-class behavior analysis technology that is more thorough and sophisticated than the older speech act theory of in-class teacher-student interaction. FIAS is made up of a matrix table for data visualization, analysis, and research goals; a coding system to characterize classroom interactions; and a set of guidelines for observing and documenting codes. The modern era of in-class grading is just getting started [19, 7].

Flipped learning is defined as when students perform tasks that are typically completed outside of class (e.g., practicing problem-solving techniques) and then return to the classroom session [5]. In contrast, traditional classroom methods—which involve presenting and transmitting information through the teaching method—are typically completed outside of the classroom and typically occur prior to class. Statistics, idea visualization, classification, clustering with associative based analysis, anomaly identification, and text-based mining are all part of the data mining approach used in the education sector [16].

Creating a new Chinese language model that defies tradition and is not constrained by time or place is a significant problem that requires immediate attention considering the advancements in multimedia and network technologies [12]. Individuals have been investigating and attempting to apply new technologies and techniques to enhance teaching and learning methods and means to increase teaching efficiency and better train abilities [2]. In addition, we seek to implement differentiated education based on each student's unique learning foundation, learning style, and other attributes, while teaching them in accordance with their aptitude [11]. But it's been hard to teach every student based on their potential because of a lack of resources for teachers and demands on teaching effectiveness.

This objective can be achieved with the help of the Intelligent Teaching System (ITS) proposal. The time and distance limits of traditional education are overcome by current Internet-based education by establishing an

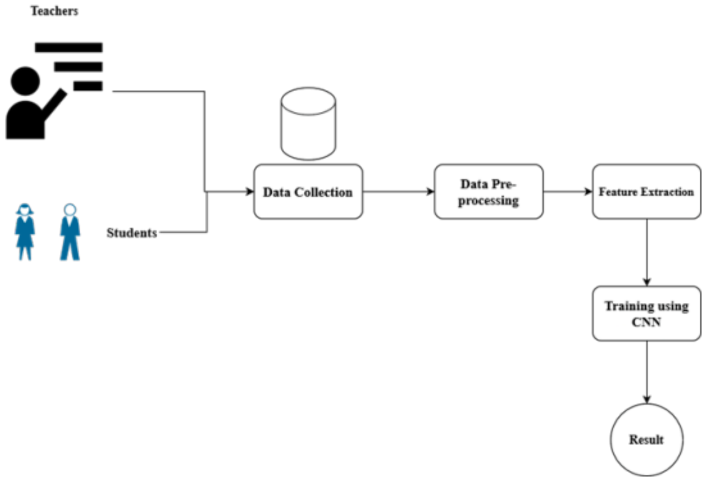


Fig. 3.1: Architecture of Proposed Method

open learning environment [1, 14]. It is essential for developing education, realizing logical resource allocation, and exploiting the resource advantages of diverse current education systems. It also provides a workable fix for the issue [10] To solve the limitations of the current network teaching system, this study develops and establishes an intelligent Chinese language network teaching system model.

Current models frequently mostly rely on textual data, including evaluations from students or the outcomes of standardized tests. The thorough integration of multimodal data sources—like audio, video, and interactive digital content—which are essential for developing a full knowledge of classroom dynamics and teacher-student interactions, is conspicuously lacking.

A lot of AI models concentrate on post-hoc examination of learning data, which reduces their usefulness for instantaneous instructional modification and real-time feedback. Real-time models that can give teachers and students immediate feedback are needed since they can greatly improve learning results.

3. Proposed Methodology. The proposed method for the Construction of teacher learning evaluation model based on CNN with equations data mining. Initially, the student feedback is collected and then the data is pre-processed. Next the data is extracted by using CNN and then the extracted features are trained by using CNN method. In figure 3.1 shows the architecture of proposed method.

The data sources are probably represented by these symbols. Instructors may offer lesson plans, instructional strategies, or assessments, while students may offer comments or performance indicators. At this point, information is gathered from the sources. It probably involves gathering a variety of data, including exam results, attendance records, student involvement, feedback, etc. The gathered data is currently being cleaned and transformed. Normalizing scores, encoding categorical variables, addressing missing values, and any other actions necessary to get the data ready for feature extraction are examples of pre-processing. Finding and extracting features from the pre-processed data that are pertinent to the learning job is the method involved in this procedure. Creating embeddings or spotting informative patterns or structures could be part of this in the context of CNNs.

A CNN model receives the features that were extracted in the preceding stage. Convolutional, pooling, and fully connected layers are the methods used by the CNN to learn from the features and modify its weights via backpropagation. The CNN model's training yielded the final output. This could be any kind of outcome that the model was trained to provide, including regression outputs like forecasting student performance or predictions like classifying teaching styles.

3.1. Data Collection and Pre-processing. Deep learning approaches are used to evaluate and interpret educational data in order to build a teacher learning evaluation model based on convolutional neural networks

(CNNs). The main elements of such a model are explained in this part, with an emphasis on the use of CNNs in the context of educational data mining for teacher assessment.

Data on education that is pertinent to the efficacy of instruction is gathered. Student assessments and feedback could be a part of this. Metrics of engagement and classroom interactions. Lesson plans and resources for teachers. Test results and grades are examples of student performance statistics. Preprocessing is done on this data to manage missing values, normalize scores, and transform qualitative input into a measurable manner. Textual data, such as feedback, is frequently transformed into numerical data using embedding and tokenization techniques.

3.2. Feature Extraction. CNNs are skilled at extracting features automatically. Word embeddings can be used to transform words into vector representations for textual input, which is subsequently fed into the CNN. Various encoding and normalizing methods are used to prepare numerical and categorical data so that a neural network can process it.

3.3. Training using CNN. These layers generate feature maps by filtering the input. Local dependencies, like word patterns in feedback or engagement data patterns, can be captured by them.

Convolutional Layers. Apply filters (kernels) to the input to create feature maps that summarize the presence of detected features in the input.

Activation Functions. Introduce non-linearities into the network, allowing it to learn complex patterns. The Rectified Linear Unit (ReLU) is a common choice for CNNs.

Pooling Layers. Reduce the dimensionality of the feature maps, making the detection of features invariant to scale and orientation.

Fully Connected Layers. These layers connect every neuron in one layer to every neuron in the next layer, making it possible to classify the image based on the features extracted by the convolutional and pooling layers.

Output Layer. Typically, a softmax activation function that converts the output of the network into probability distributions over predefined classes.

The convolution operation in the first layer can be represented as:

$$F_{ij}^{l} = \sigma(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W_{mn}^{l} \cdot I_{(i+m)(j+n)} + b^{l}) \tag{3.1}$$

where F_{ij}^{l} is the feature map at location (i,j) in layer l. W_{mn}^{l} is the weight of the kernel at position (m,n) in layer l. $I_{(i+m)(j+n)}$ is the input image or feature map from the previous layer at the corresponding location. b^{l} is the bias term for layer l. σ represents the activation function, e.g., ReLU.

Pooling (e.g., max pooling) reduces the dimensionality:

$$P_{ij}^{l} = max(Area_{ij}^{l}) \tag{3.2}$$

where P_{ij}^{l} is the output of the pooling operation at location (i,j) in layer l. $Area_{ij}^{l}$represents the area of the feature map being pooled.

The fully connected layer can be represented as:

$$O_k = \sigma(\sum_{n} W_{kn}.F_n + b_k) \tag{3.3}$$

where O_k is the output for class k. W_{kn} is the weight connecting neuron n to output k. F_n is the flattened feature map input from the last convolutional or pooling layer. b_k is the bias term for output class k.

For classification, the softmax function is commonly used:

$$Softmax(O_k) = \frac{e^{o_k}}{\sum_j e^{o_j}} \tag{3.4}$$

where e^{o_k} is the exponential of the output for class k. The denominator is the sum of exponentials of all outputs, ensuring the output is a probability distribution.

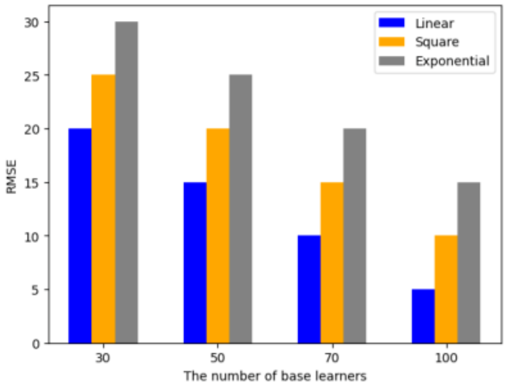


Fig. 4.1: Evaluation of RMSE based on Number of Base learners

4. Result Analysis. The information in our data set was taken from and organized within the resultant data set of four related research projects conducted by Hangzhou Hikvision Digital Technology Co., Ltd., China, our partner. These projects include: (1) ”Machine Vision-Based Video Recognition for In-Class Movement,” (2) ”Super Large- Scale Vocabulary Speech Recognition in Classrooms Situations,” (3) ”Voiceprint Recognition in Classroom Scenarios,” and (4) ”Far Field Pickup in Classroom Scenarios.” The primary data is from an actual smart learning environment implemented by Hangzhou Hikvision Digital Technology Co. There is a voice recognition pickup and two cameras in this intelligent classroom that record videos of the instructors and kids, accordingly [3].

To gather five different types of data, including student movement, student emotion, teacher movement, teacher emotion, teacher volume, and speech speed, as well as teacher speech text data for the entire class, 200 teacher samples and 300 student samples were chosen for the experiment. The audio and video data in the classroom were sampled every three seconds. The outcomes of the after-class tests are used to obtain the student label data, and the researchers’ evaluation is used to obtain the teacher label data.

One can separate the input data into two categories: sequential information, which includes student motion, feelings, quantity, and velocity, and non-sequential data, which includes teachers’ voice text. Calculating the average length of statistics and their frequency is necessary for sequence data. Furthermore, word frequency statistics must be performed for non-sequence data.

The Root Mean Square Error (RMSE) of ensemble models with varying numbers of base learners is shown in figure 4.1 as a bar chart. The discrepancy between values observed and values predicted by a model is measured by the Root Mean Square Error (RMSE). The bars are categorized into groups of 30, 50, 70, and 100 base learners based on the number of base learners in the models.

There are three sets of bars: linear, square, and exponential. These represent the various types of loss functions that are employed in statistical and machine learning models. There are four bars inside each group, each of which represents the RMSE for a particular parameter setting. The legend indicates which color corresponds to which parameter: yellow for 30, blue for 50, gray for 70, and orange for 100.These could be any hyperparameter that is adjusted during the model training process, such as learning rate, epochs, iterations, or the number of trees in a random forest. The RMSE values, a measurement of the average size of the errors between the values the model predicts and the observed values, are displayed on the y-axis. In figure 4.2 shows the result of RMSE and Loss Function.

The diagonal cells display the percentage of accurate predictions for each class, arranged from top left to bottom right. Usually, these numbers are normalized so that the total for each row is 1. In these diagonal cells, an ideal model would have 1.0 and zeros in the other cells. With a high normalized value of 0.93, ”Introduction” indicates that the model predicts this class accurately most of the time. With a rating of 0.78, ”Natural” is likewise quite high. With a diagonal value of 0.66, ”Interactive” has the lowest score, indicating that this class is the most difficult for the model to predict correctly. In the off-diagonal cells, the misclassification rates are

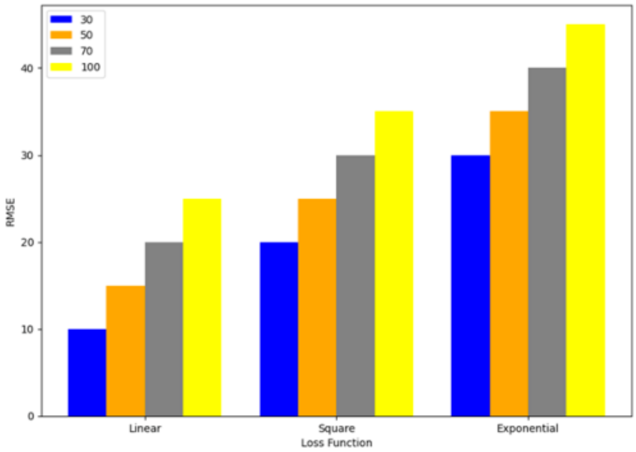


Fig. 4.2: RMSE and Loss Function

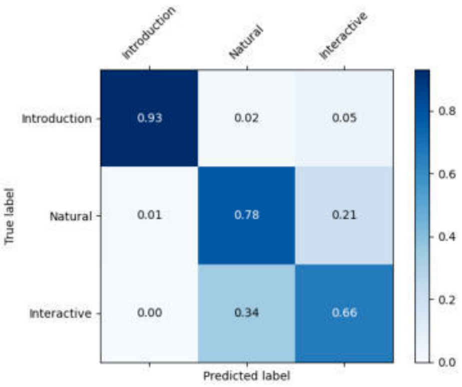


Fig. 4.3: Confusion matrix for Teachers Type

displayed. As an illustration, the word ”Introduction” is occasionally mislabeled as ”Natural” (0.02) or ”Interactive” (0.05). ’Natural’ is incorrectly categorized as ’Interactive’ (0.21) or ’Introduction’ (0.01). ’Introduction’ classes are the most accurately predicted by the model, whilst ’Interactive’ classes are less accurately predicted. ’Natural’ and ’Interactive’ classes are more frequently confused by the model with each other than ’Introduction’ with any other class. In figure 4.3 shows the confusion matrix based on teachers type.

An instrument that makes it possible to visualize an algorithm’s performance—usually supervised learning—is the confusion matrix. The projected class is represented by each column in the matrix, and the actual class is represented by each row. Predicted labels are displayed on the x-axis, while true labels are displayed on the y-axis. In this matrix, categories like ”Passionate,” ”Numerous,” and ”Solemn” have both predicted and true labels. The normalized number of observations is represented by the color shade, which ranges from 0 to 1. Higher values are generally indicated by deeper hues. The normalized count of predictions for each true/predicted label pair is represented by the integers in the matrix. In the upper left corner, for instance, the number 0.92 signifies that 92% of the actual ”Passionate” cases were accurately predicted to be ”Passionate.” In figure 4.4 shows the confusion matrix based on teachers style.

This is represented by a graph that indicates how well each model predicts the proper category for a given input. The y-axis measures this and ranges from roughly 0.65 to 0.95. The x-axis displays the following three categories (or tasks): ”Teachers’ Type,” ”Teachers’ Style,” and ”Teachers’ Media Usage.” These could stand

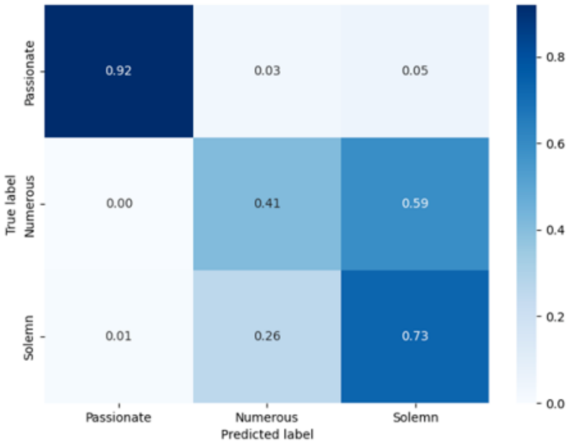


Fig. 4.4: Confusion matrix for Teachers Style

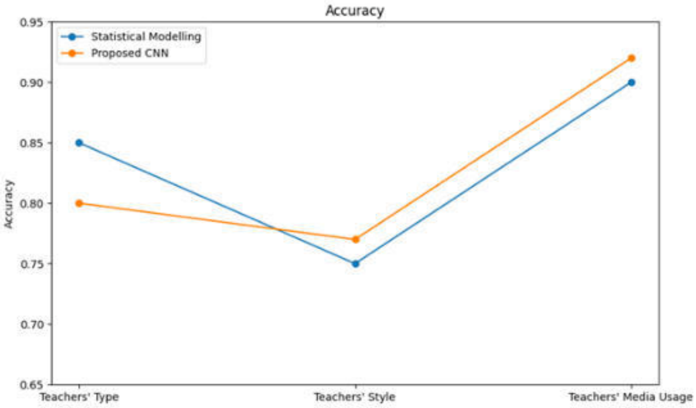


Fig. 4.5: Accuracy

in for the various facets of instruction that the models are attempting to categorize. The accuracy of each model on the three tasks is represented by the lines labeled ”Statistical Modelling” (blue) and ”Proposed CNN” (orange). The CNN marginally outperforms statistical modeling on ”Teachers' Type,” while both models perform similarly. Both models show a discernible decline in accuracy on ”Teachers' Style,” with the CNN showing a bigger decline. In figure 4.5 shows the result of Accuracy.

5. Conclusion. The importance of in-class teaching evaluation, which gauges how well students are learning in a classroom environment and how well teachers are instructing their classes, is growing as a means of tracking and improving educational standards. The concept of intelligent instruction has improved throughout time and permeated every aspect of educational application as artificial intelligence (AI) develops at a rapid rate. Because classroom instruction is so common in elementary and secondary education, the incorporation of artificial intelligence (AI) technology into the evaluation of in-class instruction has become a hotspot for research. The goals of contemporary educational systems are to increase the efficacy of instruction and personalize learning experiences for every student. In this research, we provide a novel approach that utilizes deep learning and data mining capabilities to assess teacher learning. The model's goal is to analyze and decipher the complex patterns seen in educational data in order to provide a comprehensive assessment of teacher effectiveness and student progress. In order to identify significant pedagogical indicators that are connected to successful teaching results, the model mines massive datasets, including student comments, lesson plans, classroom interactions, and performance measurements, using convolutional neural networks (CNNs). Numerous educational situations attest to the concept's efficacy, demonstrating its scalability and versatility. When used in real-world contexts, teacher assessments exhibit a discernible improvement in accuracy, providing a clear route forward for continuous educational advancement.

Funding Project. This paper is the research result of the key research of higher education teaching reform in Jilin Province in 2025, "Analysis of network learning behavior based on Data mining and its teaching strategy" (Project Number: 2025UT3S558005B).

REFERENCES

[1] S. E. Abdallah, W. M. Elmessery, M. Shams, N. Al-Sattary, A. Abohany, and M. Thabet, *Deep learning model based on resnet-50 for beef quality classification*, Inf. Sci. Lett, 12 (2023), pp. 289–297.

[2] E. Bonner, R. Lege, and E. Frazier, *Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching.*, Teaching English with Technology, 23 (2023), pp. 23–41.

[3] J. Guo, L. Bai, Z. Yu, Z. Zhao, and B. Wan, *An ai-application-oriented in-class teaching evaluation model by using statistical modeling and ensemble learning*, Sensors, 21 (2021), p. 241.

[4] B. Kang, S. Kang, et al., *Construction of chinese language teaching system model based on deep learning under the background of artificial intelligence*, Scientific Programming, 2022 (2022).

[5] T. Ley, K. Tammets, G. Pishtari, P. Chejara, R. Kasepalu, M. Khalil, M. Saar, I. Tuvi, T. Väljataga, and B. Wasson, *Towards a partnership of teachers and intelligent learning technology: A systematic literature review of model-based learning analytics*, Journal of Computer Assisted Learning, 39 (2023), pp. 1397–1417.

[6] Z. Lv, X. Wang, Z. Cheng, J. Li, H. Li, and Z. Xu, *A new approach to covid-19 data mining: A deep spatial–temporal prediction model based on tree structure for traffic revitalization index*, Data & Knowledge Engineering, 146 (2023), p. 102193.

[7] G. Mu, N. Gao, Y. Wang, and L. Dai, *A stock price prediction model based on investor sentiment and optimized deep learning*, IEEE Access, (2023).

[8] R. Ordoñez-Avila, N. S. Reyes, J. Meza, and S. Ventura, *Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review*, Heliyon, 9 (2023).

[9] S. Salem, O. Al-Habashneh, and O. Lasassmeh, *Data mining techniques for classifying and predicting teachers' performance based on their evaluation reports*, Indian Journal of Science and Technology, 14 (2021), pp. 119–130.

[10] M. Y. Shams, A. M. Elshewey, E.-S. M. El-kenawy, A. Ibrahim, F. M. Talaat, and Z. Tarek, *Water quality prediction using machine learning models based on grid search method*, Multimedia Tools and Applications, (2023), pp. 1–28.

[11] Y. Shi, F. Sun, H. Zuo, and F. Peng, *Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning*, IEEE Access, (2023).

[12] M. Shoaib, B. Shah, S. Ei-Sappagh, A. Ali, A. Ullah, F. Alenezi, T. Gechev, T. Hussain, and F. Ali, *An advanced deep learning models-based plant disease detection: A review of recent research*, Frontiers in Plant Science, 14 (2023), p. 1158933.

[13] Z. Shunxiang, Z. Aoqiang, Z. Guangli, W. Zhongliang, and L. KuanChing, *Building fake review detection model based on sentiment intensity and pu learning*, IEEE Transactions on Neural Networks and Learning Systems, (2023).

[14] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, et al., *A survey on large language model based autonomous agents*, Frontiers of Computer Science, 18 (2024), pp. 1–26.

[15] J. Xiang, *Evaluation of the college english flipped classroom teaching model based on data mining algorithms*, Mobile Information Systems, 2021 (2021), pp. 1–10.

[16] J. Xiong, T. Peng, Z. Tao, C. Zhang, S. Song, and M. S. Nazir, *A dual-scale deep learning model based on elm-bilstm and improved reptile search algorithm for wind power prediction*, Energy, 266 (2023), p. 126419.

[17] Y. Yang, C. Yu, and R. Y. Zhong, *Generalized linear model-based data analytic approach for construction equipment management*, Advanced Engineering Informatics, 55 (2023), p. 101884.

[18] L. Yuan, *Online music teaching model based on machine learning and neural network*, Soft Computing, (2023), pp. 1–12.

[19] L. Zheng, C. Wang, X. Chen, Y. Song, Z. Meng, and R. Zhang, *Evolutionary machine learning builds smart education big data platform: Data-driven higher education*, Applied Soft Computing, 136 (2023), p. 110114.

[20] W. Zheng, S. Wen, B. Lian, and Y. Nie, *Research on a sustainable teaching model based on the obe concept and the tsem framework*, Sustainability, 15 (2023), p. 5656.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments
Received: Jan 5, 2024
Accepted: Jul 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 423–431, DOI 10.12694/scpe.v26i1.3841

RESEARCH ON AUTONOMOUS NAVIGATION AND CONTROL ALGORITHM OF INTELLIGENT ROBOT BASED ON REINFORCEMENT LEARNING

YUNLONG YI*AND YING GUAN†

Abstract. The last few decades have seen impressive developments in the field of robotics, especially in the areas of autonomous navigation and control. Robust algorithms that can facilitate effective decision-making in real-time settings are needed as the need for intelligent robots that can function in complex and dynamic contexts grows. Through trial-and-error interactions with their surroundings, reinforcement learning (RL) has become a promising method for teaching intelligent agents to navigate and control robots independently. The purpose of this study is to look at the creation and use of reinforcement learning algorithms for intelligent robot control and autonomous navigation. With an emphasis on methods like deep Q-learning, policy gradients, and actor-critic approaches, the research delves into the theoretical underpinnings of reinforcement learning and how it has been applied to the field of robotics. This study assesses how well RL algorithms work to help robots acquire the best navigational strategies in challenging surroundings through an extensive literature review and empirical investigation. In addition, the study suggests new improvements and optimizations for current reinforcement learning algorithms to tackle problems unique to robot navigation, such as avoiding obstacles, routing, and interactions with dynamic environments. These improvements increase the effectiveness, flexibility, and security of independent robot navigation systems by utilizing knowledge from cognitive science and neuroscience. The suggested methods are experimentally evaluated through both real-world applications on physical robotic platforms and simulation-based research. Performance measures including navigation speed, success rate, and collision avoidance ability are used to evaluate how well the suggested algorithms operate in different scenarios and circumstances.

Key words: autonomous navigation, control algorithm, intelligent robot system, reinforcement learning

1. Introduction.

The demands of the modern logistics and warehousing industries can no longer be met by traditional manual sorting and transit efficiency due to the quick development of intelligent manufacturing and e-commerce [9]. In an assembly shop, front-line personnel can be replaced with indoor robots, enabling automation and intelligent delivery. Their automated transportation system is safer and more dependable, and their independent transit is more effective. The fast expansion of modern industry has led to increasingly complex application situations for robots [11], making the research of autonomous intelligent navigation decision-making algorithms crucial.

Several techniques and technologies must be combined to create a drone navigation system based on reinforcement learning. Sensing and understanding its surroundings: Using sensors to give the drone situational awareness is essential [5]. Drones may gather information about their environment through sensors including proximity detectors, GPS, LIDAR, and cameras. This information can be utilized to guide the aircraft and avoid obstacles. To enable continuous tracking and emergency intervention, swift and dependable connectivity is also necessary for remote control from the ground [6]. To allow the drone to make judgments depending on how it perceives its surroundings and its current condition, an effective autonomous navigation algorithm is also required.

The path planning issue has steadily grown in importance as a study topic in recent years. Conventional path planning techniques consist of the rapidly expanding random tree method [12], the artificial potential field method [2], the Dijkstra algorithm [12], the A∗algorithm, and the D∗ algorithm. But even in path planning, the conventional path planning algorithm is unable to completely comprehend the ever-more-complex and unknown external environment data. The environment's complexity makes it challenging to represent, and the prior algorithm was prone to an unstable condition of convergence. In addition, it struggles with inadequate data processing capability in large-scale areas [13]. One new sophisticated learning algorithm is called reinforcement

[*] School of Information, Shenyang Institute of Engineering, Shenyang 110136, Liaoning, China (yunlongyieanreas@outlook.com)
[†] School of Information, Shenyang Institute of Engineering, Shenyang 110136, Liaoning, China

learning.

Robotics is a field that is always changing, and autonomous navigation and control in particular has increased the need for reliable algorithms that can make decisions quickly and effectively in complex, dynamic settings. Reinforcement learning (RL) has become a powerful technique that allows intelligent agents to autonomously move and manipulate robots by means of trial-and-error interactions with their surroundings. The goal of this research is to improve reinforcement learning algorithms in order to improve their safety, flexibility, and efficiency when guiding robots through challenging situations. It seeks to investigate the integration of advanced methodologies that are essential to creating self-governing systems that are capable of learning and adapting on their own without human assistance, such as policy gradients, deep Q-learning, and actor-critic approaches.

Mobile robots' capacity to navigate autonomously is crucial because it can guarantee that the platform will get at the destination from the starting point without running into any of the many impediments in its path. Trajectory planning [7], tracking control [19], and simultaneous localization and mapping (SLAM) [1] are common steps in classical navigation techniques. But SLAM takes a long time and needs a lot of LIDAR density and precision. For mobile robots, autonomous navigation remains difficult in the absence of an obstacle map and poor range information. Consequently, scholarly interest in the unique navigation approach of end-to-end online learning based on deep reinforcement learning (DRL) has been substantial. The main contribution of the proposed method is given below:

1. We provide a novel DRL algorithm that combines the most recent methods for reward structuring, exploitation, and exploration to improve autonomous robots' decision-making and learning capabilities.
2. Our study offers a hybrid architecture that combines the best features of deliberative and reactive architectures to ensure strategic planning and real-time response in dynamic contexts.
3. We illustrate our approach's scalability by implementing it on several robotic platforms and situations. Furthermore, the learnt rules' transferability is assessed, demonstrating the algorithm's flexibility to new workloads without requiring a significant amount of retraining.
4. A thorough examination of the DRL algorithm's constituent parts is part of the study, and ablation tests are used to determine how each part contributes to overall performance, guaranteeing the results' consistency and openness.

The main research question relies on,

1. What are some ways to improve the effectiveness and adaptability of reinforcement learning algorithms in the dynamic environment of autonomous robot navigation?
2. What specific roles can cutting-edge methods like policy gradients, actor-critic methods, and deep Q-learning play in improving robot decision-making in real-time scenarios?
3. How can RL algorithms be made more contextually adaptive and efficient at learning by the integration of insights from cognitive science and neuroscience?
4. What are the shortcomings of existing reinforcement learning models when it comes to addressing the intricacies of robotic navigation in the real world, like avoiding obstacles, interacting dynamically with the surroundings, and ensuring safety?

The rest of our research article is written as follows: Section 2 discusses the related work on various autonomous navigation, control algorithm of intelligent robot,and Deep Learning Algorithms. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. Numerous fields, including home services, space exploration, automated industrial environments, and rescues, have made extensive use of robots. The main need for these applications is collision-free path planning. Consequently, path planning skills are critical to the completion of robot navigation tasks. A$*$ (A-start), RRT (rapidly explored random tree), and Dijkstra are examples of traditional path planning algorithms. To achieve the planned path, they must first comprehend all available environmental data, construct an environment model [20], and use the path-searching algorithm in accordance with predetermined optimization criteria. Since environment modeling, the basis of traditional path planning approaches, is weak and only provides local optimal solutions, it is very inaccurate when handling complicated situations.

Planning a route and identifying objects and mitigation are the two main subtasks involved in the difficult task of drone navigation. The work must be divided into smaller, optimally solved tasks because it is not always feasible to plan the entire approach ahead of time in a foreign setting [3, 10]. In unrestricted contexts where landmark placements frequently vary, simultaneous localization and mapping (SLAM) based methods are perfect. These techniques draw data from sensors such as LiDAR and IMU, although they typically add to the computing load [4]. When operating in dynamic situations, the drone's path plan needs to be revised on a regular basis to account for impediments that are identified while in flight and come into its path.

Afterwards, intelligent bionic path planning techniques with some autonomies were developed; these mostly consisted of particle swarm optimization [14], optimizing ant colonies [18], and genetic algorithms [21]. The intelligent bionic algorithm can perform planning path tasks in a dynamic space; however, when the computational load is high, path planning efficiency is low and real-time path planning efficiency is not ensured [8]. Furthermore, the intended path is not the best option when the robot does not know enough about its working surroundings. These conventional path planning methods typically have trouble digesting highly dimensional, complex data on the environment in challenging situations, or they are prone to local optimal performance.

Tasks involving path planning are often solved by adapting artificial intelligence techniques. It is simple to use brute-force or exhaustive search methods for UAV path planning jobs [15]. Although they can be quite slow, the breadth-first and depth-first search for space techniques are thorough and always locate a path if one is available, or the shortest of all accessible paths. To prevent dead ends, they can also be used in conjunction with backtracking [16]. Although they can also be used to speed up search, greedy techniques always run the danger of piling into neighbourhood minima. Because more than 50 targets make brute-force search impractical, these approaches begin with one or more quick fixes and work their way up to the best answer by applying local modifications and, if necessary, random restarts [17].

Although reinforcement learning (RL) has shown great promise for robotic control, current algorithms frequently encounter difficulties related to the dynamic and unpredictable nature of real-world situations. Managing high-dimensional sensory inputs, learning optimal policies quickly, and reacting in real time to environmental changes are some of these issues. Moreover, a lot of training data and computer power are needed for most RL techniques, and they could not scale well in real-world applications or generalize well in other contexts. The development of RL algorithms that can dependably function in a variety of operational contexts without sacrificing performance or safety, interpret complex sensory data fast, and adapt to new situations are still far from being fully developed.

3. Proposed Methodology. The proposed methodology for autonomous navigation and control algorithm of intelligent robot is evaluated by using Deep Reinforcement Method (DRL). At first, build a virtual world in which the robot can function. This could be a fully virtual environment created to test scenarios, or it could be a digital duplicate of an actual location. Next, based on the specifications of the problem, select an appropriate Deep Reinforcement Learning (DRL) algorithm, such as Deep Q-Networks (DQN). To better meet the navigation and control problems unique to the intelligent robot, train the selected DRL algorithm. Create a system of rewards that incentivizes the desired behavior. Assessing the robot's ability to navigate and control itself in real-world settings and gathering performance statistics. Return the design, adjusting the robot control systems and DRL model in response to evaluation results. In figure 3.1 shows the architecture of proposed method.

The indoor robot's autonomous navigation system is a noteworthy technological development in robotic control, providing several technical advantages essential for dependable and effective operations in changing interior environments. The robot can navigate difficult spaces with great accuracy thanks to the integration of GPS-INS (Global Positioning System-Inertial Navigation System) assistance, AHRS (Attitude and Heading Reference Systems), and high-precision dynamic 3D processing. This degree of accuracy is necessary for operations like material handling and distribution in warehouses and manufacturing plants, which call for precise movement and placement.

Perpetual motion formulas, which take into account ongoing changes in position and velocity, provide the foundation for the robot's sophisticated navigational abilities. This allows the robot to move smoothly and respond quickly to changes in its surroundings. The robot's radar system, which gathers vital information such as target spectrum, bearings, and velocity to enable reliable collision avoidance and accurate docking procedures, significantly improves these capabilities. In addition to helping with navigation, this radar technology enables

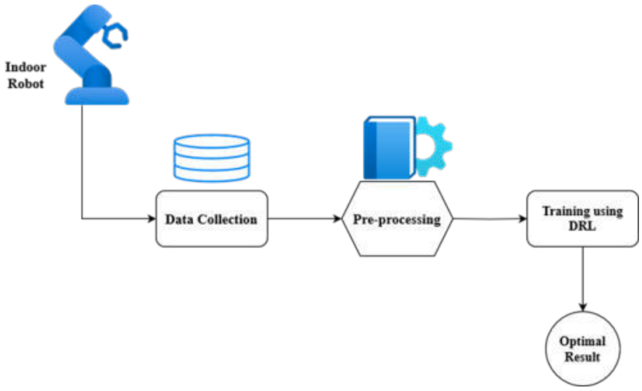


Fig. 3.1: Architecture of Proposed Method

the robot to keep an eye on its surroundings and take proactive measures to avoid potential obstructions or modify its course as needed.

Moreover, the robot's intelligent control mechanism heavily relies on its array of distance sensors. These sensors continuously scan the surroundings to offer real-time information on the location and size of impediments, allowing the robot to quickly alter its course. In highly populated or unpredictable areas, this sensor-based method to obstacle identification is essential for protecting the robot's operational integrity and guaranteeing safety.

3.1. Indoor Robot for Intelligent Control Mechanism. The indoor robot has sophisticated automated navigational capabilities, including high-precision dynamic 3D processing data, AHRS, and GPS-INS inertial navigation support systems, to fulfill the distribution duty. The robot's perpetual motion formulas are determined using the following equations to guarantee brevity and universality in the 2D plane:

$$\begin{bmatrix} \dot{x}\left(t\right) \\ \dot{y}\left(t\right) \\ \dot{\varphi}\left(t\right) \\ \dot{v}\left(t\right) \end{bmatrix} = \begin{bmatrix} v(t)\cos\phi(t) \\ v(t)\sin\phi(t) \\ \omega\left(t\right) \\ a(t) \end{bmatrix} \tag{3.1}$$

If ϕ denotes the robot's motion guidance, v is its velocity, and x and y stand for the robot's 2D coordinates in relation to its surroundings. The equations that follow can be used to characterize the status report of time t during the interval:

$$\begin{cases} x\left(t\right) = x\left(t-1\right) + v\left(t-1\right)\Delta t\cos\phi\left(t-1\right) \\ y\left(t\right) = y\left(t-1\right) + v\left(t-1\right)\Delta t\sin\phi\left(t-1\right) \\ v\left(t\right) = v\left(t-1\right) + a\left(t-1\right)\Delta t \\ \phi\left(t\right) = \phi\left(t-1\right) + \omega(t-1)\Delta t \end{cases} \tag{3.2}$$

The indoor robot has a radar attached to pick up signals. Target spectrum, bearings, velocity, and other data can be obtained by comparing the broadcast signal with the received target echo. This gives fundamental information for navigating, avoiding collisions, parking spaces, and other tasks. The relative azimuth angle φ and the distance in relation D among the indoor robots and the point of interest can be determined at any time t. Furthermore, assume that the goal position vector S_d and the robot position vector S_{uav} are

$$S^{uav} = [x_t, y_t, z_t]^T \; S_d = [x_t, y_t, z_t]^T \tag{3.3}$$

Furthermore, the robot's observation equation to the target point in a two-dimensional coordinate system at a specific time is defined as

$$z = \begin{bmatrix} D \\ \varphi \end{bmatrix} = \begin{bmatrix} ||\left(x_t, y_t\right) - \left(x_d, y_d\right)||_2 \\ \text{actan}\frac{y_t - y_d}{x_t - x_d} \end{bmatrix} \tag{3.4}$$

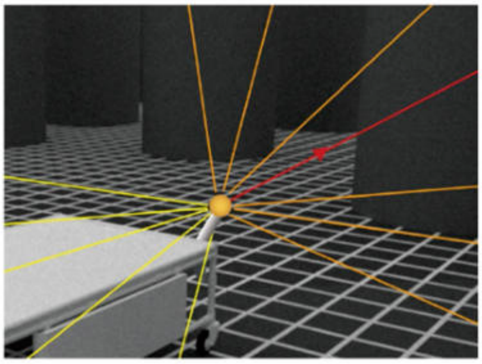


Fig. 3.2: Structure of Agent Sensor

The primary obstacle to autonomous navigation and intelligent control of the robot is a complex and dynamic environment. The robot must identify environmental hazards to navigate autonomously. Therefore, the robot is equipped with a dozen distance sensors to aid in its detection of any obstructions that may be within its detectable range across the front. The robot's ability to detect obstacles at any given time is defined as follows:

$$O_o = [d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}] \tag{3.5}$$

wherein the data sent by the respective sensor is represented by the numbers $d_1, d_2, \ldots, d_3$. The sensor information in front of the agent is represented by the red line in Figure 3.2, the sensor signal in advance of the agent (which includes the robot's right and left sides) is represented by the orange-colored lines, and the sensor signal coming from behind the agent is represented by the yellow lines. If the sensor finds no obstacles, its maximum detectable range is set to L; if it detects obstacles, the distance between the agent and the obstacle is represented by $d_n \in [0, L]$.

3.2. Autonomous Navigation for Indoor Robot using DRL. Large-scale and complicated situations are challenging to handle using the classic path planning algorithm due to its poor convergence speed and high processing requirements. Deep reinforcement learning can be implemented in an end-to-end observation and management systems with strong flexibility through the combination of the perceptual capacity of deep learning with the decision-making power of reinforcement learning. This can significantly increase the effectiveness of path planning.

The agent engages with the surroundings at each instant to acquire a high-dimensional observation, and the deep learning approach may be used to perceive the state attributes. The action's value function is assessed by considering the anticipated return, and the action that corresponds to the current state is mapped to it. By repeatedly performing the procedures, the environment reacts to this action and receives the subsequent observation, allowing the best possible approach to be determined.

In particular, the Markov decision process, symbolized by a quadruple (S, A, R, c), can be used to show the entire process of learning of an agent. The agent's observations and state are represented by the quad S; the tasks that the agent can perform are represented by A; the reward function, R, represents the agent's rewards upon completion of an action in a particular state; and the discount coefficient, c, balances immediate and accumulative rewards during the learning process. In figure 3.3 shows the structure of DRL.

4. Result Analysis. Simulation tests are put up to confirm the efficacy of the DRL algorithm in smart navigation and autonomous control of indoor robots. The Gym-agent-master system, Python 3.6, TensorFlow 1.14.0, and PyCharm are used to execute the environment that has been simulated. The cylinder in the simulated environment is a barrier, and it is generated in the Northeast geodetic coordinate system using the VTK third-party software. The barriers in the scenario being simulated have a radius of one meter and a centre-to-centre distance of three meters. The robot's maximum running speed is set at 2.0 m/s. To guarantee the

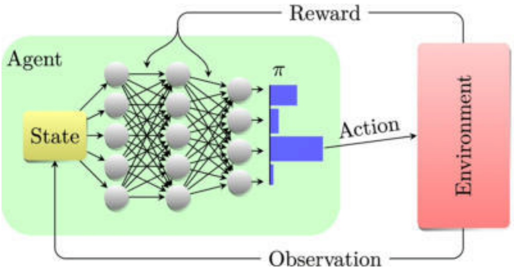


Fig. 3.3: Structure of DRL

efficacy of the navigational task, the robot and destination must be at least 50 meters apart at the beginning, and the simulated phase should last for one second.

In the test of simulation, a simulator is built to enable the robot to handle itself intelligently and autonomously in expansive, complicated settings. The robot's dynamical physical restrictions are disregarded, and its shape is abstracted as a sphere to facilitate the experiment of simulation. To guarantee scarce prizes, there must be a minimum separation of more than 30 meters between the starting point and the destination. The robot has a rangefinder attached so it can be observed. The goals of self-navigating and intelligent operation are considered accomplished when the robot approaches the destination and the distance between it and its intended location is less than one meter.

The AGV's training process begins after the pertinent parameters are specified. This round will be considered ended if the robot cannot finish the training assignment or encounters a barrier within the allotted time. After ten, the experiment will restart, and the subsequent round will start. Simulating the real world, the situation's update rules are established as follows: the robot's setting, its destination, and the number of barriers in every round are all randomly determined.

The AGV's training process begins after the pertinent parameters are specified. This round will be considered ended if the robot cannot finish the training task or collides with an obstacle within the allotted time. At ten, the game will restart and the subsequent round will start. Simulating the real world, the scenario's update rules are established as follows: the robot's setting, its final destination, and the amount of obstacles in every round are all arbitrarily determined.

In the simulation experiment, the robot is trained using the DRL, DDPG, and TD3 algorithms, respectively, to confirm the effectiveness of the proposed DRL algorithm in automatic navigation and intelligent control. As seen in Figure 4.1, the robot's reward value at the end of each training round is recorded.

The DRL algorithm exhibits the most pronounced increasing trend, as seen in Figure 4.2, and it takes the lead to reach the high of 240 after roughly 4000 rounds. The TD3 algorithm exhibits severe fluctuations and the lowest return performance. With a high fluctuation, the classic DDPG method does not begin to rise until around 2000 rounds, and it peaks later than the optimized DRL algorithm. However, the DRL algorithm's reward value declined erratically after about 6600 training episodes. However, after about 7100 rounds, it swiftly recovered to a higher, steady level. This demonstrates how the DRL algorithm suggested in this research might enhance training effects by assisting the robot in adapting to the noisy training environment.

The achievement rate for 0–10000 rounds during the training of the suggested DRL, DDPG, and TD3 algorithms is displayed in Figure 4.2. It is evident that under the DDPG and TD3 computations, the robot's job completion rate is less than 80%, and its learnt methods perform poorly. The DRL algorithm training success rate has the highest growing trend in comparing. The success rate is consistently above 80% after 3000 rounds, with a peak value approaching 90%. When compared to the other two algorithms, the DRL algorithm offers the best learning method and the highest success rate.

We conducted 1000 rounds of comparative tests in each of the three scenarios mentioned above to confirm the effectiveness of the robot autonomous navigation strategy under the DRL algorithm. The success rates of indoor robot navigation are displayed in Table 4.1.

In the testing procedure, we simultaneously logged the data of every successful round and calculated the

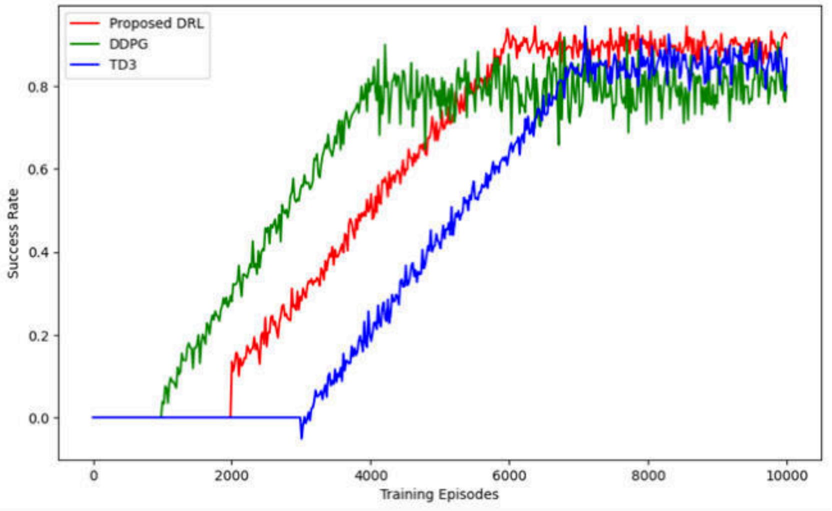


Fig. 4.1: Evaluation of Success rate

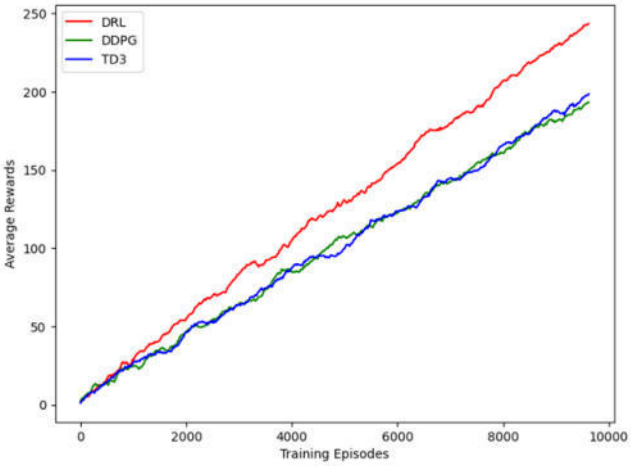


Fig. 4.2: Success rate of Completion of Robots

Table 4.1: Success Rate of Robot capturing autonomous navigation

Methods Used	**100 obstacles**	**150 obstacles**	**200 obstacles**
Proposed DRL	92.35%	83.905%	77%
DDPG	80%	65%	61.32%
TD3	85%	72%	65%

average job completion time under each of the three algorithms, as indicated in Table 4.1. The three algorithms do not significantly differ in how long navigation tasks take in simple settings.

Following training, the intelligent system of control will be evaluated in three different contexts to confirm the efficacy of the indoor robot system navigation approach. The environmental barriers are numbered 100, 150, and 200, in that order. Figure 4.3 displays the outcomes of the simulation. Based on the simulation experiment

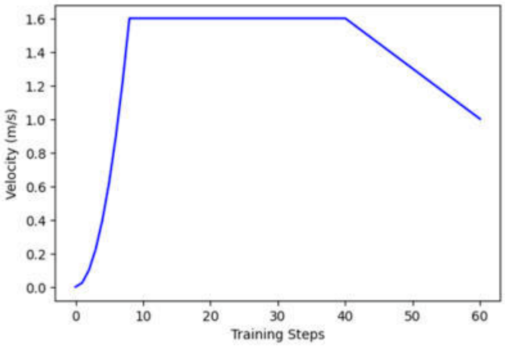


Fig. 4.3: Evaluation of Training steps and Velocity of Robot

results, it may be inferred that an experienced robot is capable of intelligent self-navigating in an environment with varying numbers of obstacles, allowing it to avoid obstacles and reach its target. Additionally, the robot can raise its speed gradually and keep it inside the limit of maximum speed until it reaches its destination, based on the trend of robot speed change.

From the first step to about step 10, the velocity climbs rapidly until it reaches a plateau. From step 10 until about step 40, the plateau has a constant speed just over 1.4 m/s. The velocity gradually decreases after step 40 and stays that way until step 60, the last step shown. This kind of graph could be used to illustrate a simulation or experiment in which the velocity of a vehicle or robot is tracked over time or via repeated training phases. The plateau is a time when the velocity is sustained at its highest level. There are several possible explanations for the decrease: the introduction of a deceleration protocol, the commencement of a limiting factor (such as energy depletion), or changes in the surroundings that could impact velocity.

5. Conclusion. In summary, the study of reinforcement learning-based autonomous navigation and control algorithms for intelligent robots marks a substantial breakthrough in the discipline of robotics. Intelligent robots may learn and change their navigation and control methods in changing circumstances with explicitly programming them by utilizing reinforcement learning methods. The effectiveness and adaptability of algorithms that use reinforcement learning in empowering robots to move around and carry out activities in complex and unexpected environments have been proven by this study. Robots can effectively investigate their surroundings, pick up knowledge from encounters, and gradually improve their decision-making abilities by using reward signals to direct learning. Additionally, the research's conclusions have ramifications for several practical uses, such as automation in industries, service robotics, and driverless cars. Intelligent robots' capacity to navigate and adjust to shifting conditions on their own offers the potential to improve production, safety, and efficiency in a variety of settings. To fully realize the potential for intelligent machines in ever-more complex and dynamic environments, further study and development in this field will be necessary to enhance and maximize self-navigating and control computations, solve issues with adaptability and generalizations, and more.

Subsequent investigations could concentrate on creating increasingly complex sensor fusion algorithms that more successfully combine data from diverse sources including radar, LiDAR, and visual cameras. The robot's ability to see and make decisions in congested or dynamically changing settings may be enhanced by this integration. Navigational judgments could be greatly improved by incorporating machine learning algorithms that use historical data to forecast future environmental situations. Deep learning techniques for predicting possible impediments and human movement patterns in indoor environments could be investigated further.

6. Project information. Basic research project of Liaoning Provincial Department of Education in 2023: Design and research of intelligent positioning and control system for quadruped robots based on big data clustering (JYTMS20230321).

REFERENCES

[1] K. Almazrouei, I. Kamel, and T. Rabie, *Dynamic obstacle avoidance and path planning through reinforcement learning*, Applied Sciences, 13 (2023), p. 8174.

[2] M. Caruso, E. Regolin, F. J. Camerota Verdù, S. A. Russo, L. Bortolussi, and S. Seriani, *Robot navigation in crowded environments: A reinforcement learning approach*, Machines, 11 (2023), p. 268.

[3] X. Chen, S. Liu, J. Zhao, H. Wu, J. Xian, and J. Montewka, *Autonomous port management based agv path planning and optimization via an ensemble reinforcement learning framework*, Ocean & Coastal Management, 251 (2024), p. 107087.

[4] V. D. Cong et al., *Path following and avoiding obstacle for mobile robot under dynamic environments using reinforcement learning*, Journal of Robotics and Control (JRC), 4 (2023), pp. 157–164.

[5] J. Escobar-Naranjo, G. Caiza, P. Ayala, E. Jordan, C. A. Garcia, and M. V. Garcia, *Autonomous navigation of robots: Optimization with dqn*, Applied Sciences, 13 (2023), p. 7202.

[6] W. Fei, Z. Xiaoping, Z. Zhou, and T. Yang, *Deep-reinforcement-learning-based uav autonomous navigation and collision avoidance in unknown environments*, Chinese Journal of Aeronautics, 37 (2024), pp. 237–257.

[7] W. Huang, Y. Zhou, X. He, and C. Lv, *Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation*, IEEE Transactions on Intelligent Transportation Systems, (2023).

[8] S.-L. Jeng and C. Chiang, *End-to-end autonomous navigation based on deep reinforcement learning with a survival penalty function*, Sensors, 23 (2023), p. 8651.

[9] W. Li, M. Yue, J. Shangguan, and Y. Jin, *Navigation of mobile robots based on deep reinforcement learning: Reward function optimization and knowledge transfer*, International Journal of Control, Automation and Systems, 21 (2023), pp. 563–574.

[10] D. Ma, X. Chen, W. Ma, H. Zheng, and F. Qu, *Neural network model-based reinforcement learning control for auv 3-d path following*, IEEE Transactions on Intelligent Vehicles, (2023).

[11] E. E. Montero, H. Mutahira, N. Pico, and M. S. Muhammad, *Dynamic warning zone and a short-distance goal for autonomous robot navigation using deep reinforcement learning*, Complex & Intelligent Systems, 10 (2024), pp. 1149–1166.

[12] S. Na, T. Rouček, J. Ulrich, J. Pikman, T. Krajník, B. Lennox, and F. Arvin, *Federated reinforcement learning for collective navigation of robotic swarms*, IEEE Transactions on cognitive and developmental systems, 15 (2023), pp. 2122–2131.

[13] F. Naseer, M. N. Khan, and A. Altalbe, *Intelligent time delay control of telepresence robots using novel deep reinforcement learning algorithm to interact with patients*, Applied Sciences, 13 (2023), p. 2462.

[14] J. E. Sierra-Garcia and M. Santos, *Combining reinforcement learning and conventional control to improve automatic guided vehicles tracking of complex trajectories*, Expert Systems, 41 (2024), p. e13076.

[15] Q. Sun, L. Zhang, H. Yu, W. Zhang, Y. Mei, and H. Xiong, *Hierarchical reinforcement learning for dynamic autonomous vehicle navigation at intelligent intersections*, in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 4852–4861.

[16] Z. Sun, Y. Fan, and G. Wang, *An intelligent algorithm for usvs collision avoidance based on deep reinforcement learning approach with navigation characteristics*, Journal of Marine Science and Engineering, 11 (2023), p. 812.

[17] T. Wang, V. Dhiman, and N. Atanasov, *Inverse reinforcement learning for autonomous navigation via differentiable semantic mapping and planning*, Autonomous Robots, 47 (2023), pp. 809–830.

[18] Z. Xu, B. Liu, X. Xiao, A. Nair, and P. Stone, *Benchmarking reinforcement learning techniques for autonomous navigation*, in 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 9224–9230.

[19] Y. Xue and W. Chen, *Multi-agent deep reinforcement learning for uavs navigation in unknown complex environment*, IEEE Transactions on Intelligent Vehicles, (2023).

[20] C. Yan, G. Chen, Y. Li, F. Sun, and Y. Wu, *Immune deep reinforcement learning-based path planning for mobile robot in unknown environment*, Applied Soft Computing, 145 (2023), p. 110601.

[21] Y. Yin, Z. Chen, G. Liu, and J. Guo, *A mapless local path planning approach using deep reinforcement learning framework*, Sensors, 23 (2023), p. 2036.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making
in Large-Scale Environments
Received: Jan 5, 2024
Accepted: Jul 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 432–440, DOI 10.12694/scpe.v26i1.3844

RESEARCH ON PERSONALIZED LEARNING RECOMMENDATION SYSTEM BASED ON MACHINE LEARNING ALGORITHM

SIQI LI[*]AND DEMING LI[†]

Abstract. The educational system has started to implement more individualized material from conventional functions in recent years due to the ongoing advancements and developments in science and technology, particularly the continuing growth of artificial intelligence, machine algorithms, and other technologies. The standardized approach to teaching used by conventional educational institutions frequently ignores the individual requirements and learning preferences of every student. To improve learning outcomes, a system of education that is personalized and enhanced by algorithms using machine learning can offer individualized learning materials and suggestions that reflect every student's educational background, interests, and skills. Additionally, machine learning methods may offer immediate feedback on student achievement and modify instructional strategies in response to that feedback. Making sure AI is employed to promote higher education's overarching objectives, like encouraging creativity and critical thinking, as opposed to merely eliminating chores and boosting effectiveness, is another difficulty. This paper examines over the several ways that artificial intelligence (AI) and the Optimized Collaborative Filtering Algorithm are being used in higher education. It also proposes an approach for increasing students' cognitive abilities and compares it with different methods that are currently in use. It has been demonstrated that, in comparison to other models, the suggested model performs better by achieving 95% recall and 99% testing accuracy.

Key words: Personalized learning, recommendation system, machine learning, students, educational institutions

1. Introduction.

The rise of technological advances in educational institutions is leading to innovative teaching and learning, with artificial intelligence and intelligent instruction taking the lead. This trend is expected as education becomes more and more digitized [1, 10, 5]. The time of artificial intelligence is confronted with a new challenge as big data in education grows and needs to be analysed to ensure correct forecasting. Education-related large data analysis and forecasting can be satisfied by machine learning, a significant area of artificial intelligence. The suitableness of machine learning and smart instruction are thus explored via an examination of the method, the thing, particular methodologies, and users of machine learning.

There are difficulties and moral issues with using AI in higher education. Maintaining the impartiality and precision of AI systems while eliminating possible prejudices is one of the major difficulties. Challenges exist around the confidentiality of student information in addition to the possibility that AI will eventually take the role of human educators and assistance personnel [20, 13]. Making sure AI is applied in a manner that advances the general objectives of higher learning, like encouraging innovative thinking and problem-solving, as opposed to only becoming used for task automation and productivity gains, is a further challenge [6].

With the widespread use of the latest iterations of computer networking, wireless communication, cloud computing, the Internet of Things (IoT), and artificial intelligence (AI), we are entering an entirely novel phase of smart technological advances and extensive data mining uses. All facets of our life are being controlled by artificial intelligence, machine learning, and big data processing, which are driving global digital transformation [12, 2]. New teaching and learning approaches have been implemented in higher education institutions because of these groundbreaking advancements in information technology.

The expanding breakthroughs in artificial intelligence and machine learning, in particular, have made the incorporation of personalized learning in educational systems increasingly important. Traditional teaching methods frequently produce less than ideal results since they do not take into account each student's particular demands and learning preferences. In order to provide an extremely flexible learning environment, this

[*]International Academy of Arts, Jilin International Studies University, Changchun 130117, Jilin, China
[†]School of Education, Jilin International Studies University, Changchun130117, Jilin, China (Corresponding author, lideming@jisu.edu.cn)

study suggests a novel application of an Optimized Collaborative Filtering Algorithm along with other AI-driven techniques. These technologies allow for the real-time assessment and modification of teaching tactics based on immediate feedback on student performance. They also provide personalized learning materials and recommendations depending on each student's academic background, interests, and skills.

Nevertheless, there is still work to be done in utilizing AI to improve more general educational objectives like encouraging creativity and critical thinking, in addition to efficiency. This study examines the use of AI in higher education in a variety of contexts, offering a novel strategy to improve cognitive capacities and providing empirical support for its superiority over current models. The purpose of this study is to demonstrate how contemporary AI implementations can overcome conventional constraints and provide notable gains in educational efficacy and customisation.

This study investigates a variety of machine learning approaches, such as the Optimized Collaborative Filtering Algorithm, for recommending appropriate models of recommendation for use in higher education settings. These models assist students in identifying their true focus and ability. Additionally, it enhances the kids' cognitive capacities and attitudes in general. The main contribution of the proposed method is given below:

1. An innovative AI algorithm is presented to enhance the learning capacity of students in higher education.
2. The suggested algorithm makes it easier to find, read, and grasp a text piece in large data quickly and in real time.
3. The suggested algorithm stresses pupils' quality and self-worth while simultaneously enhancing their cognitive abilities.
4. In addition to outperforming these models in many ways, it is demonstrated that the suggested algorithm performs better when compared to other common reference models, such as the Gram-CF, CNN-CF, and CNN recommendation models.

The rest of our research article is written as follows: Section 2 discusses the related work on various personalised recommendation system and Machine Learning Algorithms. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. Even while artificial intelligence has advanced significantly in the present, it is still in the early stages of integrating with learning and instructional methods and is still far from a mature state. Investigations on the creation and utilization of technological advances has gained a lot of interest recently due to the cases and study findings of the current coupling of artificial intelligence with educational systems both domestically and internationally. Researchers have focused particularly on the four main uses of artificial intelligence in higher education, which are "intelligent learning guide structure, automatic evaluation system, educational games, and educational robots" [8, 3, 19, 7]. Additionally, it is the primary approach taken by educators working in fundamental education.

Numerous methodologies have been suggested in the field of individualized learning route generating study. Personalized learning path generating techniques based on deep learning have also drawn a lot of interest [4]. The neural collective filtering method is a popular recommendations technique that effectively handles limited data as well as cold start issues by embedding knowledge on users and products using neural networks. Scholars have also been experimenting with using neural collective filtering methods to solve path generation problems in customized educational path creation [9]. Neural collective filtering techniques, for instance, are employed to create individualized learning paths by predicting learners' interests and level of knowledge mastery [15].

Additionally, creating a personalized learning path makes extensive use of cognitive diagnostic tools [16]. Cognitive diagnostic methods seek to determine and assess the cognitive level of learners to assist them in identifying their learning deficiencies and problems and to offer appropriate educational recommendations and assistance [11]. By combining cognitive diagnostic approaches with neural collaborative filtering computations, it is possible to generate individualized learning paths by more effectively recommending points of knowledge and learning materials that align with learners' interests and cognitive levels [14].

Choosing into account the significant influence on utilizing the Internet, the abundance of chances for utilizing new skills and more complex behavioural patterns, and the scientific foundation that artificial intelligence

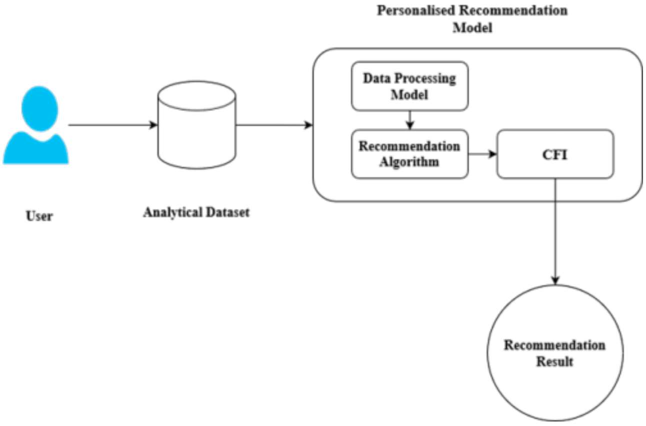


Fig. 3.1: Architecture of Proposed Method

contributes to, it improves students' university education by expanding the growth of their educational requirements [18]. Additionally, AI, communication, and big data support the real-time provision of constructive cultural and educational suggestions to users, that can help satisfy university students' expanding needs for growth [17]. Conversely, the use of AI in educational institutions gives schools and universities the chance to build amicable relationships with their students and can, in certain cases, enhance their sense of identity and self-worth.

3. Proposed Methodology. The proposed methodology for Personalized learning recommendation system based on machine learning algorithm based on artificial intelligence (AI) and the Optimized Collaborative Filtering Algorithm. Initially, the higher education dataset is collected and then the data is pre-processes. Next the Personalized learning recommendation system is trained by using Optimized Collaborative Filtering (CFI) Algorithm. In figure 3.1 shows the architecture of proposed method.

3.1. Construction of Personalized Learning Recommendation System. The rise in popularity of online learning has raised the bar for intelligent resource recommendation, with students, instructors, and administrators serving as the system's user objects. The latter involves an evaluation of users and resources, whereas the first two are primarily concerned with individualized choice of resources and recommendations for instruction. The present suggested model still has certain problems with algorithm stability and accuracy, though, and this need be investigated further and improved.

3.2. Optimized CFI Method. The optimization algorithm chosen in this study is based on the CFI System (CF), which enables the model to perform intelligent recommendations. The secret to this algorithm's success is its ability to compute resource and user similarities, as well as classify and limit the data to find possible user growth regions. As a result, the system can manage unstructured resource data and does away with the need for resource feature modeling. As seen in Figure 3.2, there are two types of CF algorithms: user-based and project-based.

The method will begin with the resource that the user prefers, look for items that are comparable to it, eliminate the items which have already been ensued, and then offer suggestions. Users benefit from greater freshness with the former and relative stability with the latter. Based on various practical objectives, the model should select suitable recommendation principal methods. This algorithm's benefits include the how users and feedback information interact, how easy it is to approach, and how stability improves with time. Still, this approach has a high degree of user interaction; for example, the model's calculation accuracy will drastically decrease if the user's behaviour data is poor.

This suggests that there are three main processes that make up the system's functioning: First, features are extracted and categorized using both structured and unstructured data. Formula (3.1) indicates that structural

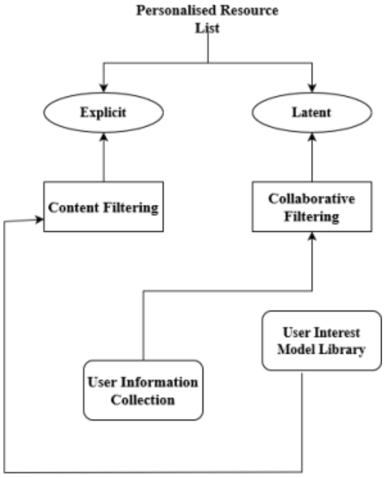

Personalised Resource List
Explicit
Latent
Content Filtering
Collaborative Filtering
User Interest Model Library
User Information Collection

Fig. 3.2: Structure of CFI

transformation is necessary for unstructured data.

$$\begin{cases} D = \{d_1, d_2, \ldots, d_N\} \\ T = \{t_1, t_2, \ldots, t_n\} \\ d_j = \{w_{1j}, w_{2j}, \ldots, w_{nj}\} \end{cases} \tag{3.1}$$

Using the items selected as a model, the collection of articles is represented by D in equation (3.1) above. T stands for a set of items that include certain keywords. d$_j$ is the set of vectors that make up text. The total amount of items and words are denoted by N and n, accordingly. W$_{nj}$ is a representation of each keyword's value. The weight calculation method chosen by the research is the phrase recurrence inverse document frequency technique, as indicated by formula (3.2).

$$TF - IDF\left(t_k, d_j\right) = TF\left(t_k, d_j\right).log\frac{N}{n_k} \tag{3.2}$$

The average number of recurrence of the k-th word in article j is represented by the expression $TF(t_k, d_j)$ in equation (3.2) above. The total amount of items in the set with k words is denoted by n_k. Formula (3.3) thus displays the relative importance of the term in item j.

$$w_{kj} = \frac{TF - IDF\left(t_k, d_j\right)}{\sqrt{\sum_{s=1}^{T} TF - IDF(t_s, d_j)^2}} \tag{3.3}$$

CFI methods are used in this study to investigate consumers' possible interests. It is difficult to identify users having comparable preferences because multiple descriptions and other details regarding the same resources may exist in the framework. Based on this rationale, this study presents a method for calculating similarities that combines behavior and content. It is represented as a rating of similarity $sim_{grade}(u, v)$ and content similarity, respectively. Formula displays the result of the former computation.

$$sim_{grade}\left(u, v\right) = \frac{\sum i \in D_u \cap D_v \frac{1}{\log(1+|U(i)|)}}{\sqrt{|D_u|\,|D_v|}} \tag{3.4}$$

The resource's assessment variety of users u and v is denoted by D_u and D_v in equation (3.4) above, respectively. The user group that has left comments on resource d_i is represented by U (i). Formula (3.5) displays the $sim_{content}(u,v)$ content similarity between two users.

$$sim_{content}\left(u,v\right)=\frac{EM_u.EM_v}{\left|EM_u\right|.\left|EM_v\right|} \tag{3.5}$$

The data sets of two users' initial interest are represented by EM_u and EM_v, accordingly, in equation (3.5) previously. In conclusion, formula (3.6) displays the results of the mixed similarity calculation.

$$sim\left(u,v\right)=\beta sim_{grade}\left(u,v\right)+(1-\beta)sim_{content}(u,v) \tag{3.6}$$

The weighting factor, denoted by β in equation (3.6) above, is a similar ratios variable that must be empirically determined within the interval [0, 1]. When the amount of weighting is zero, the model is sufficient to take similarities in content into account. On the other hand, the model simply must take score similarity into account if the factor to be weighted is 1. To determine the ultimate user's mixed similarity, the algorithm must first compute the score similarity and content similarity among users independently. The similarity values are then fused based on the weighting factor values.

A person who most closely resembles the user who is being targeted will be added to the neighbouring user collection, and CFI concepts will then be applied to suggest potentially interesting resources. Formulas illustrate that a feature word f$_i$'s weight is determined in the latent preferences model.

$$w2_{uj}=\sum v_i\in U_M\frac{sim\left(u,v_i\right)}{\sum v_i\in U_u sim\left(u,v_i\right)}.w1v_{ij} \tag{3.7}$$

An efficacy test using a model simulation was carried out prior to as well as following optimization to confirm the efficacy of the optimized CFI recommendation model. Additionally, the suggested model was used on the real system for visual evaluation, revealing how many clicks and successful suggestions the model received in each week. In the meantime, the efficacy of the optimized customized recommendations model was contrasted with that of other recommendation models.

The model described in this study has a technological edge since it makes sophisticated use of the Optimized Collaborative Filtering Algorithm (CFI), which improves the model's capacity to provide individualized learning experiences. In dynamic situations such as e-commerce, traditional collaborative filtering systems frequently suffer from sparse data and scalability issues. The model used in this study, however, solves these difficulties with the help of clever recommendation algorithms that, in addition to calculating similarities between users and resources based on their interactions, also efficiently classify and filter the data to identify possible growth areas for user interest.

Through the integration of user- and project-based collaborative filtering techniques, the model is able to provide suggestions that strike a compromise between stability and freshness, accommodating user preferences while upholding the integrity of the recommendation framework. The model's input is further refined by using sophisticated data preprocessing techniques, such as TF-IDF for feature extraction from unstructured data, which makes recommendations that are more pertinent and accurate. The model is able to fine-tune its predictions through the final mixed similarity calculation, which combines behavioral and content similarities with empirically established weights. This provides a reliable solution that can be customized to unique user needs and behaviors.

Essentially, the technical complexity of the model not only overcomes the shortcomings of conventional collaborative filtering when handling sparse and complex data, but it also improves the capacity to provide customized recommendations in an educational setting, which could revolutionize the way that personalized education is approached in digital platforms.

4. Result Analysis. To address the overfitting issue, some neurons must be randomly removed. According to a review of the literature, the test and training loss curves level out, and the amount of loss drops below 0.06, indicating that the model has converged, when the number of iterations approaches 40. Accuracy can be significantly improved when compared to the standard recommendation's models CNN, FCNN-CF, and Gram-CF.

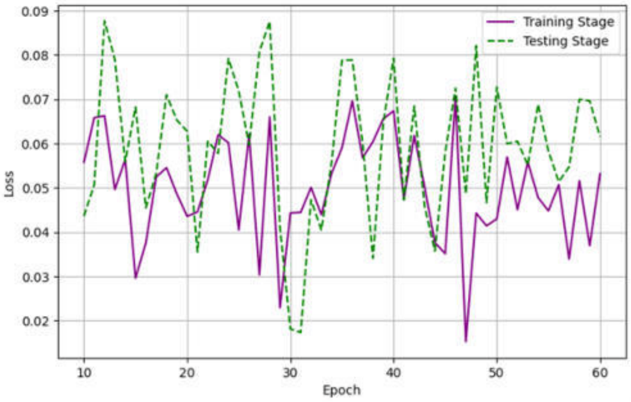


Fig. 4.1: Training and Testing Loss

To validate the model, same ecological and assessment criteria are employed. The standard method CNN bases itself on users' run information, the algorithm FCNN-CF depends on user conduct statistics, and the Gram-based revised recommendations engine Gram-CF are selected as reference model.

The F1 value, precision, accuracy, and recall rate are the evaluation metrics that are employed. Nonetheless, when contrasted with the original models, the suggested model exhibits improvements in both recall and precision. In the meantime, F1 has significantly improved. As a result, these findings demonstrate that the suggested model outperforms the CNN, FCNN-CF, and Gram-CF comparison approaches in every way. This enhancement is the result of two-dimensional feature extraction, which combines temporal and spatial information to strengthen the expressive of the features. This eliminates anomalous data and enhances the model's capacity for learning.

The first few epochs see a sharp decline in training and testing losses, indicating that the model is picking up patterns from the data quickly. Both lines exhibit some convergence following the first decline, suggesting that the pace of learning has leveled off. This is a typical stage where the model keeps becoming better. The model may not be substantially overfitting to the training data based on the reasonably close distance between the training and testing lines. Overfitting would be indicated if the training loss was significantly less than the testing loss. The variations may suggest that the model's learning is not totally stable, particularly in the testing line. In figure 4.1 shows the training and testing loss.

With each epoch, the model is learning substantially from the data, as evidenced by the high growth in both lines at the beginning. when the epochs increase, both accuracies level out, which is typical when the model gets closer to its maximum ability to learn from the available data. Usually, when both lines converge and the testing accuracy closely tracks the training accuracy, the model is not overfitting. A high training accuracy but a significantly lower testing accuracy would be indicative of overfitting. A peak in testing accuracy appears to be reached below 1.00, which may indicate that the model has performed as well as it can in terms of generalization given the present design and data. In figure 4.2 shows the training and testing accuracy.

A recommendation system based on fully convolutional neural networks with collaborative filtering (CF) is represented by this line. Recall declines as the quantity of items rises, according to the pattern. This is an example of a convolutional neural network-based recommendation algorithm. As more things are taken into consideration, the recall drops, much like the FCNN-CF. This probably refers to an approach that combines collaborative filtering with a gram matrix or feature representation. This is the suggested algorithm from the study's author(s), which is where this chart came from. It displays the recall percentage for their method, which likewise decreases with an increase in the number of objects. As there are more goods, the recall gets worse. In figure 4.3 shows the result of Recall.

This illustrates the accuracy of a collaborative filtering system with a fully convolutional neural network for varying item counts. It appears to function steadily across the range. indicative of the precision performance

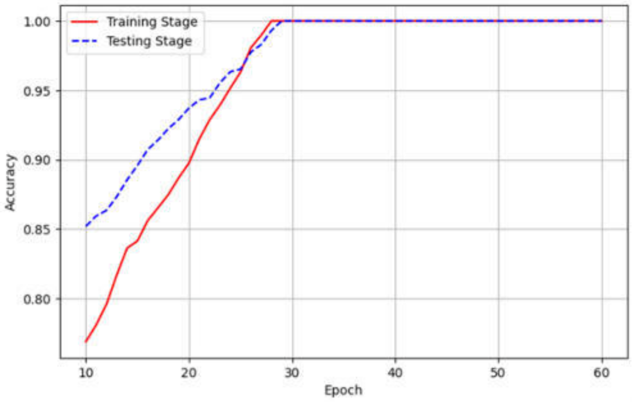


Fig. 4.2: Training and Testing Accuracy

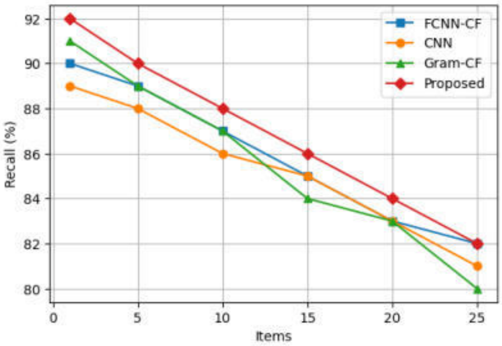


Fig. 4.3: Recall

of a conventional convolutional neural network. It exhibits a decrease as the number of items rises, much like the others. This is probably an approach that combines Collaborative Filtering with a gram matrix or some other equivalent feature representation. As additional items are added, its precision declines, which is a typical pattern in recommendation systems. This line corresponds to the accuracy of the algorithm that the graph's designers have suggested. It displays a performance comparison with the other algorithms. In figure 4.4 shows the result of Precision.

5. Conclusion. The continuous progress in the fields of science and technology, especially the development of artificial intelligence, machine learning, and other methods, has led educational institutions to begin implementing more customized content from conventional uses in the past few years. Traditional colleges and universities usually disregard each student's unique needs and educational preferences in Favor of a standardized teaching style. A personalized learning system powered by machine learning algorithms can provide tailored learning materials and recommendations based on each student's educational history, interests, and abilities, thereby improving learning outcomes. Furthermore, machine learning techniques have the potential to provide instantaneous feedback on student accomplishment and adjust educational strategies accordingly. Another challenge is ensuring that AI is used to support the broad goals of higher education, such as fostering creativity and critical thinking, rather than just removing tasks and increasing efficiency. This study looks at the various applications of the Optimized Collaborative Filtering Algorithm and artificial intelligence (AI) in higher education. Additionally, it suggests for enhancing pupils' cognitive capacities and contrasts it with various approaches that are now in use. It is shown that the proposed model performs better than the other models.

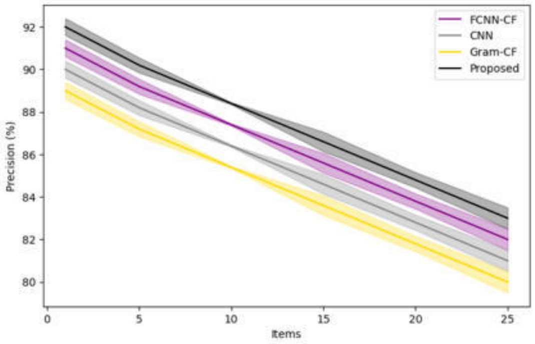


Fig. 4.4: Precision

In order to develop immersive learning environments, future research could investigate how to integrate the personalized recommendation system with AR and VR technology. This could potentially increase engagement and retention rates by enabling learners to interact more interactively with the content.

Funding Project. This paper is the research result of the key research of higher education teaching reform in Jilin Province in 2025, "Analysis of network learning behavior based on Data mining and its teaching strategy" (Project Number: 2025UT3S558005B).

REFERENCES

[1] A. Al Ka'bi, *Proposed artificial intelligence algorithm and deep learning techniques for development of higher education*, International Journal of Intelligent Networks, 4 (2023), pp. 68–73.

[2] A. Alper, *A resource recommendation for improving musical expression and narration in piano education: An examination of loeschhorn op. 65 etudes.*, Educational Research and Reviews, 16 (2021), pp. 189–201.

[3] F. E. Alsaadi, Z. Wang, N. S. Alharbi, Y. Liu, and N. D. Alotaibi, *A new framework for collaborative filtering with p-moment-based similarity measure: Algorithm, optimization and application*, Knowledge-based systems, 248 (2022), p. 108874.

[4] W. Bao, H. Che, and J. Zhang, *Will_go at semeval-2020 task 3: An accurate model for predicting the (graded) effect of context in word similarity based on bert*, in Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 301–306.

[5] W. Chen, Z. Shen, Y. Pan, K. Tan, and C. Wang, *Applying machine learning algorithm to optimize personalized education recommendation system*, Journal of Theory and Practice of Engineering Science, 4 (2024), pp. 101–108.

[6] I.-A. Chounta, E. Bardone, A. Raudsep, and M. Pedaste, *Exploring teachers' perceptions of artificial intelligence as a tool to support their practice in estonian k-12 education*, International Journal of Artificial Intelligence in Education, 32 (2022), pp. 725–755.

[7] Z. Cui, P. Zhao, Z. Hu, X. Cai, W. Zhang, and J. Chen, *An improved matrix factorization based model for many-objective optimization recommendation*, Information Sciences, 579 (2021), pp. 1–14.

[8] Y. Dai and J. Xu, *Study of online learning resource recommendation based on improved bp neural network*, International Journal of Embedded Systems, 14 (2021), pp. 101–107.

[9] S. Du, L. Li, Y. Wang, Y. Liu, and Y. Pan, *Application of hpv-16 in liquid-based thin layer cytology of host genetic lesions based on ai diagnostic technology presentation of liquid*, Journal of Theory and Practice of Engineering Science, 3 (2023), pp. 1–6.

[10] X. Feixiang, *Intelligent personalized recommendation method based on optimized collaborative filtering algorithm in primary and secondary education resource system*, IEEE Access, (2024).

[11] Z. He, X. Shen, Y. Zhou, and Y. Wang, *Application of k-means clustering based on artificial intelligence in gene statistics of biological information engineering.*

[12] C. Jung and L. Mertins, *Quality and accountability in the online business education environment*, Computer, 54 (2021), pp. 49–52.

[13] S. C. KONG, H. OGATA, J.-L. SHIH, and G. BISWAS, *The role of artificial intelligence in stem education*, in Proceedings of the 29th International Conference on Computers in Education Conference, 2021.

[14] L. Pan, J. Xu, W. Wan, and Q. Zeng, *Combine deep learning and artificial intelligence to optimize the application path of digital image processing technology*, (2024).

[15] Y. Pan, Z. Shen, Y. He, K. Wei, and Y. Zhang, *Application of three-dimensional coding network in screening and diagnosis of cervical precancerous lesions*, Frontiers in Computing and Intelligent Systems, 6 (2023), pp. 61–64.

[16] T. Song, Q. Zhang, G. Cai, M. Cai, and J. Qian, *Development of machine learning and artificial intelligence in toxic pathology*, Frontiers in Computing and Intelligent Systems, 6 (2023), pp. 137–141.

[17] W. Sun, L. Pan, J. Xu, W. Wan, and Y. Wang, *Automatic driving lane change safety prediction model based on lstm*, arXiv preprint arXiv:2403.06993, (2024).

[18] W. Wan, W. Sun, Q. Zeng, L. Pan, and J. Xu, *Progress in artificial intelligence applications based on the combination of self-driven sensors and deep learning*, in 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE), IEEE, 2024, pp. 279–284.

[19] X. Xiong, X. Li, Y. Hu, Y. Wu, and J. Yin, *Handling information loss of graph convolutional networks in collaborative filtering*, Information Systems, 109 (2022), p. 102051.

[20] L. Yang, H. Wang, J. Zheng, X. Duan, and Q. Cheng, *Research and application of visual object recognition system based on deep learning and neural morphological computation*, International Journal of Computer Science and Information Technology, 2 (2024), pp. 10–17.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making
in Large-Scale Environments
Received: Jan 5, 2024
Accepted: Jul 6, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 441–449, DOI 10.12694/scpe.v26i1.3845

USING GENETIC ALGORITHM TO OPTIMIZE THE TRAINING PLAN AND GAME STRATEGY OF BASKETBALL PLAYERS

QIU DAWEI*

Abstract. Basketball teams rely heavily on the effectiveness of their practice regimens and game plans to succeed. The intricacy of the game and the range of skills players must possess make it extremely difficult to develop effective training plans and tactical techniques. The use of genetic algorithms (GAs) as a cutting-edge technique to enhance basketball players' practice regimens and game strategies is investigated in this work. Inspired by biology and natural selection, genetic algorithms provide a potent optimization method that, via the repeated steps of evolution, can find close to ideal solutions to challenging issues. As chromosomes made up of genes relating to different parameters and tactics, prospective training plans and game tactics are represented by GAs, which enable them to efficiently search across the large solution space and find combinations that optimize desired results. As part of the research approach, the goals and limitations of the optimization issue are defined. Fitness functions are intended to assess each potential solution's efficacy, directing development toward better solutions across a series of iterations. This study shows how effective genetic algorithms are in optimizing basketball players' training regimens and game strategy through simulators and real-world tests. In order to continually enhance player growth and team competitiveness, coaches can use GAs to refine and modify tactics based on feedback as well as performance data in an iterative manner. The results of this study have significance for team sports other than basketball, where results are heavily influenced by the interaction of players' individual abilities, teamwork, and decisions about strategy. In the end, incorporating algorithms based on genetics into sports analytics presents a viable way to improve coaching techniques and reach the highest levels of efficiency in sporting settings.

Key words: genetic algorithm, optimization, game strategy, basketball players, deep learning

1. Introduction. The development of technology for computer vision has led to a growing emphasis on technological achievements, especially in sports [7]. Videos can be used for human action identification and behavioral analysis, but this equipment can also help officials make better decisions about sports motions and give players, trainers, or analysts insights into correct motion strategies. In actuality, the field of sports science research is booming these days, with the goal of tracking players, identifying them, and identifying the activity they execute. Therefore, increasing the total efficacy of athletic training requires making coaches' training plans more scientific in character and utilizing cutting-edge technologies like computer vision to analyse athletes' performances [8, 20, 2].

An essential part of teaching and fostering pupils' healthy development is basketball. Participating in basketball games can help children develop their strong will, unwavering spirit, camaraderie, and coordination skills in addition to their endurance [21]. Basketball instruction is still taught using the conventional method of having instructors explain concepts, provide examples, and have students mimic them mechanically. This approach necessitates a high degree of independence for students and interaction between teachers and pupils. The total impact and caliber of basketball instruction are significantly diminished by this full-house irrigation-based teaching strategy, which also stifles pupils' capacity for original thought and creativity [6]. This is the final goal of teaching physical education, and it is a crucial issue that educators must address. Basketball is a contemporary sporting event as well as a full exercise game.

Basketball instructors design customized training schedules for players to improve their abilities. In the past, coaches made training strategy decisions based on their own experiences as well as the athlete's technical proficiency [1]. Evaluating this approach's success is challenging because it heavily relies on personal bias and necessitates a thorough examination of motion. Efficiency and precision are essential components of training in modern sports. Teachers can greatly enhance training results if they can correctly determine an athlete's

*Department of Physical Education, Guangxi University of Traditional Chinese Medicine; Nanning, Guangxi; 530200; China (qiudaweispace1@hotmail.com)

sports posture [13]. Therefore, identifying sports postures precisely is essential to developing training regimens that make scientific sense and improving the performance of athletes. The combination of estimation of human poses and action recognition systems has been made possible by improvements in basketball sports, and this has a significant impact on raising the points scored rate [4].

Teams must constantly adapt and improve their practice plans and game plans to keep ahead of the intense demands of competitive basketball. Making efficient training plans and in-game strategies is difficult due to the intricacy of synchronizing the talents and tactics of several players. Conventional approaches might not be able to adequately convey the dynamic character of the game and the interaction between different tactical components. Innovative strategies that may methodically and effectively assess and enhance team performance are required in this circumstance. Inspired by biological evolution, genetic algorithms (GAs) provide a potent optimization tool that can search large solution spaces and find optimal techniques that traditional methods might miss. The potential for GAs to transform basketball strategy and training by adjusting to real-time performance

The main contribution of the proposed method is given below:

1. To ensure continuous progress and optimal performance, our system uses a genetic algorithm to customize training routines that adjust to the changing skill sets and fitness levels of basketball players.
2. We created a dynamic framework for game strategy that use evolutionary algorithms to construct intricate in-game strategies, optimizing plays by taking opponent data and real-time player statistics into account.
3. We present a thorough analysis of the differences between our GA-optimized technique and conventional training and strategy methods, highlighting the superior results in terms of player growth, team performance, and strategy adaptation.

Research helps to identify solutions for

- In comparison to conventional coaching techniques, how good are genetic algorithms in optimizing basketball training schedules and game strategies?
- What are the best parameters and strategies to reflect basketball training and strategy optimization that may be efficiently stored as genes within the chromosomes of genetic algorithms?
- In a genetic algorithm framework, what are the best fitness functions to assess basketball practice plans and tactics?

The rest of our research article is written as follows: Section 2 discusses the related work on various Basketball player, game strategy, sports teaching and Deep Learning Algorithms. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. Among the more popular sports in the globe is basketball. This societal impact also corresponds with a high level of academic documentation, as this is one of the sports with the most research output [10]. These studies focus on a variety of topics, including game indicators (GI), psychological factors, nutritional elements, tactics and techs, and physical fitness (PF) [14, 15, 5, 9]. Some research methods are used independently of one another, with no connection among the various subjects. Basketball is a complicated sport with many interacting variables; hence the progress of basketball study involves multidisciplinary studies including numerous objects of study. By examining any potential connections among physical fitness and game indicators, this study seeks to determine the relationship between the two of them.

Furthermore, GI may be impacted by several ambient factors. In accordance with the classification [22], three categories of matches—equal, unbalanced, and extremely imbalanced—were specified in this line based on the outcomes. Based on this categorization and the analysis [11] concluded that whereas two-point shooting determines the outcome of lopsided games, the quantity of rebounds determines even games. Furthermore, the researcher [18] verified that the beginning and substitute players' contributions to the team and their GI differed. The authors of [12] claimed that players displayed variations in their physical and technical-tactical characteristics based on the game situation. These specifications had to do with what every position in the game did.

Even though the previously discussed deep learning-based algorithms can identify basketball position, they frequently exhibit weak points when faced with complex contextual factors as shifting illumination, crowded

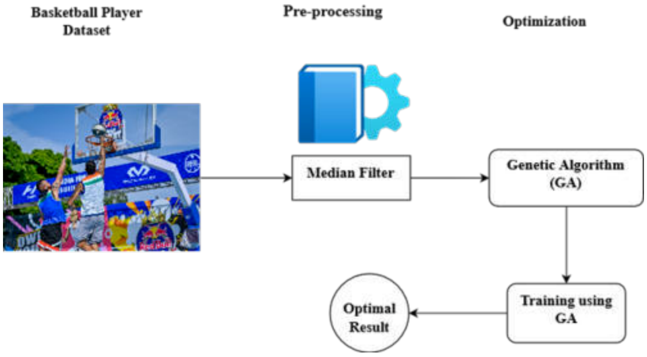


Fig. 3.1: Architecture of Proposed Method

backgrounds, and perspective alterations [19]. Furthermore, the basic design of the conventional C3D system makes it challenging to accurately detect basketball stance. Its effectiveness is therefore low, and the error rate is considerable. Furthermore, most of these methods require evaluating the entire movie or applying a classifier to each frame to assign just one action name to the entire video [3]. These methods are less effective than the human vision tactics, which can identify a scene from only one instance of visual input.

Considering these circumstances, this work investigates the use of cluster analysis and the association rule's method in the instruction of basketball in a mobile computing surroundings [16]. It does so in a way that will effectively encourage the use of these techniques in college physical education and serve as a guide for the improvement of physical learning in China. Because of the database's current security measures, users' interactions with the database are typical in real-world application settings. The standard access records for each user role are those kept in the database log. Various user jobs have quite diverse access behaviors [17].

3. Proposed Methodology. The proposed method for optimizing the plan and game strategy of basketball players using Genetic Algorithms (GA). Initially, the basketball players images or video is collected and then the data is pre-processed by using Median Filter. Next the GA method is used for optimizing the variables. Finally, the optimized variables are trained and predicted. In figure 1 shows the architecture of proposed method.

Predictive models are then trained using the variables that the GA has optimized. These models are made to predict the results of various training approaches, giving coaches and trainers the ability to evaluate possible effects prior to implementing training methods in full. Lastly, the training process incorporates the improved variables, allowing for the monitoring of their efficacy and the validation of predictions against real performance results. This cyclical cycle makes it possible to continuously improve tactics in light of actual performance, enabling a dynamic, data-driven, and adaptive approach to sports training.

3.1. Basketball Data Collection and Analysis. Basketball action identification is significant as it can assess player efficiency and give coaches and players feedback on areas that require development. It can also be utilized to help officials make better calls throughout games by examining the actions of players on the court to recognize fouls or other infractions that might have gone unnoticed. The extreme diversity in player motions and actions presents one of the biggest obstacles to action recognition in the game of basketball. Basketball matches feature numerous players running all over the court at once, which makes it challenging to follow each player precisely and identify their movements. Furthermore, a lot of basketball moves could appear similar yet, depending on the situation, have distinct meanings.

3.2. Pre-processing. Aside from the preprocessing procedure, the proposed model consists of two primary sections. During the data preparation process, basketball photos may have a significant amount of noise.

In this research, a median filter is applied as a stage of preprocessing to address this problem and reduce interference when identifying basketball activity from video pictures.

3.3. Optimization and Training using Genetic Algorithms. A search heuristic known as a Genetic Algorithm (GA) was developed in response to Charles Darwin's notion of natural evolution. It reflects the process of natural selection, in which the most fit individuals are chosen to procreate and give rise to the subsequent generations of humans. The goal of using a genetic algorithm to optimize the method of assessment for the Basketball Player analysis is to identify and select the best teaching techniques and achievements.

Encoding. Every person would be encoded as a chromosome, which is commonly shown as an array of numbers or a binary string. Every gene on a chromosome would stand for a choice variable, like whether a player gets chosen.

$$Chromosomes = (gene_1, gene_2, \ldots\ldots, gene_n) \tag{3.1}$$

3.2.2 Initialization

Create a starting population (individuals) of possible solutions. This population may be created at random, using heuristics or previous knowledge as a basis.

$$Pop = \{Chromosome_1, Chromosome_2, \ldots\ldots, Chromosome_n\} \tag{3.2}$$

Fitness Function. Create a fitness function that measures each player's (combination of players') performance according to the specified parameters. This function could consider several variables in the context of basketball player forecasting, including player roles, team science, opponent strengths, and player statistics (points scored, rebounds, contributions, etc.).

$$Fitness = (Chromosome_i) = f(gene_1, gene_2, \ldots\ldots, gene_n) \tag{3.3}$$

Selection. To produce the future generation, choose certain people from the existing population. People are usually chosen through a fitness-based procedure in which those who perform better are more likely to be chosen. Rank-based, tournament, and roulette wheel selection are examples of popular selection techniques.

$$Sel = Selection(Population) \tag{3.4}$$

Crossover. Select individuals should undergo crossover (recombination) to produce offspring for the following generation. To generate novel solutions, this entails the exchange of genetic material, or genes, between pairs of humans. The problem and the solution representation determine the crossover technique that is employed.

$$Child_1, Child_2 = Crossover(Parent_1, Parent_2) \tag{3.5}$$

Mutation. To preserve diversity and investigate new areas of the search space, introduce random modifications, or mutations, to some members of the population. The algorithm is kept from becoming trapped in local optima by mutation. Mutations can occur when bits in binary strings are switched around or when chromosomal gene values are slightly altered.

$$Mutated\_Child = Mutation(Child) \tag{3.6}$$

Replacement. For several generations or until a termination condition (such as a maximum number of generations or convergence requirements) is satisfied, repeat the selection, crossover, and mutation processes.

$$Population_{new} = Replacement(Ppopulation, Offspring) \tag{3.7}$$

The phases 3 through 7 of the GA process must be repeated iteratively to develop the solutions and get the highest fitness score.

This research makes it possible to create highly individualized training plans that are suited to the unique strengths and limitations of both the team and individual players by using GAs. Improved player development and more successful training results may result from this individualized strategy.

GAs offer a framework for continuously updating and improving strategies in response to performance data. This process is known as dynamic strategy adaptation. This flexibility keeps the squad competitive by guaranteeing that workout routines and game strategies work against diverse opponents and in variable settings.

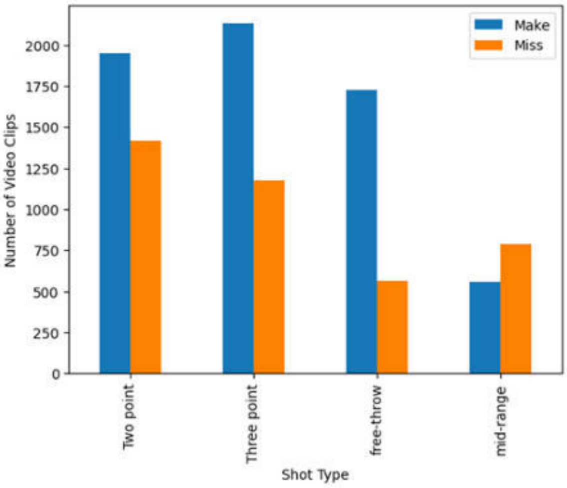


Fig. 4.1: Samples in each class on the Basketball-51 dataset

4. Result Analysis. The implementation of the suggested approach was carried out using two distinct datasets, Basketball-51 and SpaceJam, to verify its effectiveness. Two components comprise the basketball player activities found in the SpaceJam dataset. The joint locations of the single-player posture are shown in samples in the first part, which is referred to as the joint dataset. Every frame in RGB colours for every action is contained in the second portion, which is referred to as the clip database. Actions like walk, not doing anything, run, protection, dribbling, ball in hand, pass, block and select, and shoot are all included in this dataset. The completed datasets with annotations contain roughly 32,560 samples, which can be used as the basis for the training and testing stages.

The first dataset, referred to as the ”clip Dataset,” consists of 16 RGB frames that highlight a single participant in each case. The joint coordinates (x,y) of the participant on the picture plane are contained in the second dataset, referred to as the ”joint Dataset,” on the opposite side. Although the file extensions of the two datasets are different—joints are saved as numpy vector files (name file.npy), and clips are stored in compressed mp4 files (name file.mp4)—they share the same identifier for related examples.

The 10,311 video snippets from 51 NBA basketball matches make up the Basketball-51 collection. Third-party recording devices, commonly employed in sports broadcasting, recorded all the videos. The video footage was initially divided into eight class designations: mid-range shot miss, mid-range shot makes, free throw fail, free throws make, two-point miss, and three-point miss. Figure 2 shows the arrangement of data in the dataset according to several labels. the quantity of video clips, divided into categories based on the many basketballs shot kinds, including mid-range, free throw, three-point, and two-point. There are two bars for each type of shot, one for each shot that was made and one for missed. Shots that were made (”Make”) are represented by blue bars, and shots that were missed (”Miss”) are represented by orange bars. The number of successful (made) versus unsuccessful (missed) attempts for each type of shot can be compared. Although the actual numbers are not displayed in this description, such a chart usually enables readers to quickly compare these values.

When an algorithm for deep learning is being trained, the epochs stand for iterations across a dataset. We have epochs numbered 0 to just over 50 on the x-axis. Accuracy is represented by the y-axis, which runs from 0 to 1 (or 0% to 100%). Both lines begin with accuracy near zero, and both models get more accurate as the number of epochs rises. This is typical of machine learning models, which improve their accuracy in classifications or predictions as they ’learn’ from more data across more epochs. Early in training, the ”proposed model” line surpasses the YOLOv4 line and continues to perform better for the remaining epochs. It appears that both models’ accuracy peaks around the 50th epoch, suggesting that further training after this may not yield appreciable gains in accuracy. In figure 3 shows the evaluation of training accuracy.

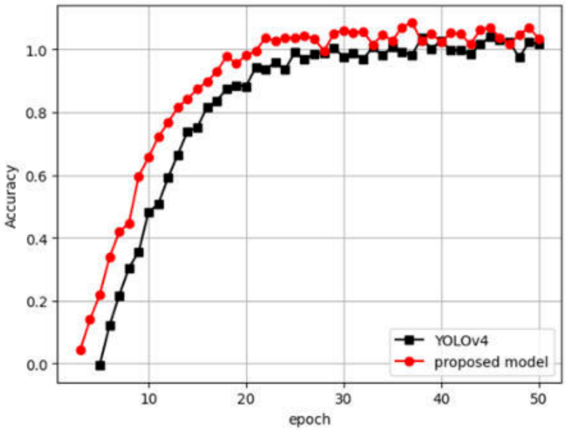


Fig. 4.2: Evaluation of Training Accuracy

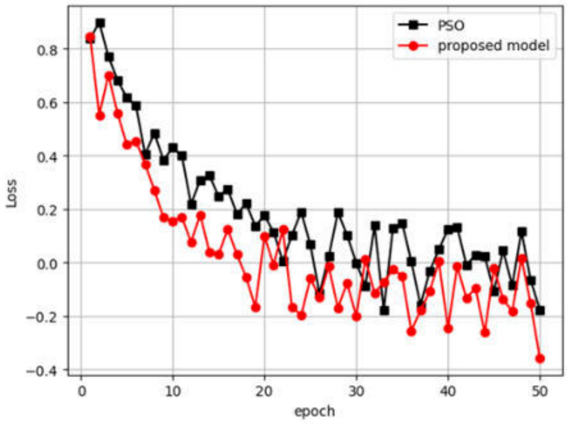


Fig. 4.3: Evaluation of Training Loss

The epoch number, which ranges from 0 to little over 50, is indicated on the x-axis. The error between the values that the model predicts and the actual values from the training data is typically shown by the y-axis, which measures loss. During machine learning training, minimizing this loss is usually the aim. Both models have an initial larger loss (about 0.6 for the proposed model and closer to 0.8 for PSO), which generally decreases as the number of epochs rises. This suggests that the models are learning and developing.

The suggested model, whose loss curve displays more notable ups and downs, is the only one on the graph that displays significant changes in the loss. This can point to fluctuations in the training procedure or noise in the training set. Eventually, the loss values of both models approach or fall below zero, with the suggested model sporadically falling below zero. This could indicate overfitting or a mistake in the loss computation. In figure 4 shows the evaluation of Training Loss.

In Figure 5,6 shows the confusion matrices of the suggested model for subject-dependent and independent classification techniques based on recollection and classification reports. Most misclassifications are found to occur in the mid-range area, where shots are typically labeled as two- or three-point attempts. This might be explained by the imbalance in the dataset, which has more 2- and 3-point data clips than mid-range ones. Lower performance may also be caused by the same similarity of interclass activities and the absence of a clear range differentiation between various ranges. The suggested approach much outperforms previous groups in

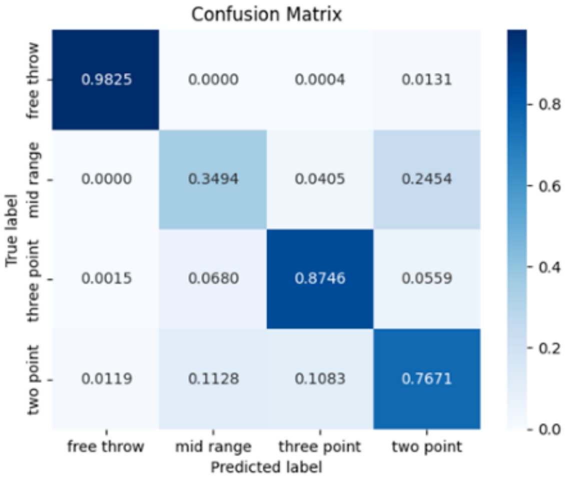


Fig. 4.4: Confusion Matrix for Subject Dependent Basketball-51 dataset

identifying free throw characteristics with high accuracy.

The Predicted Labels are represented by the x-axis, and the True Labels are represented by the y-axis. These labels match the categories that the model is predicting, which in this example are the many kinds of basketball shots that can be made free throw, mid-range, three-point, and two-point. The heatmap's right side features a color scale that shows the range of values inside the matrix: 1.0 (red) denotes a higher frequency of predictions, while 0.0 (blue) denotes a lower frequency. Every cell in the heatmap represents the likelihood (or frequency) of the model's predictions.

At the intersection of the 'free throw' (genuine Label) and 'free throw' (Predicted Label), there is a dark red cell with a 0.9549 value, suggesting a high frequency of accurate predictions in cases when the genuine shot turned out to be a free throw. On the other hand, a low-value cell is coloured blue, suggesting that the model seldom misclassifies a two-point shot as a free throw. An example of this would be the cell at the intersection of "two point" (True Label) and "free throw" (Predicted Label). The diagonal cells of a perfect confusion matrix, which run from top left to bottom right, should have the greatest values (darker red), signifying that the model correctly predicts the actual class labels. Misclassifications are shown by off-diagonal cells with greater values. In figure 6 shows the confusion matrix for Subject Independent Basketball-51 dataset.

5. Conclusion. The success of basketball teams is largely dependent on how well their practice routines and game strategies work. It is quite challenging to create efficient training regimens and tactical strategies due to the complexity of the game and the variety of talents players need to possess. This paper investigates how basketball players might improve their practice routines and game strategy by using genetic algorithms (GAs), a cutting-edge technique. Genetic algorithms are powerful optimization techniques that, via the iterative processes of evolution, can find almost perfect solutions to difficult problems. These algorithms draw inspiration from biology and natural selection. Prospective training strategies and game tactics are represented by GAs, which are chromosomes composed of genes pertaining to various parameters and tactics. This allows GAs to search the vast solution space effectively and identify combinations that maximize desired outcomes. The objectives of the optimization problem and its constraints are specified as part of the research methodology. Fitness functions are designed to evaluate the effectiveness of each possible solution, guiding development toward more effective solutions through a series of iterations. Using simulators and actual testing, this study demonstrates the efficacy of genetic algorithms in optimizing basketball players' training plans and game strategies. Coaches can use GAs to iteratively adapt and adjust tactics based on feedback and performance data, ultimately enhancing player growth and team competitiveness. The study's findings are relevant to team sports other than basketball, where players' individual skills, teamwork, and strategic decisions all interact to greatly impact

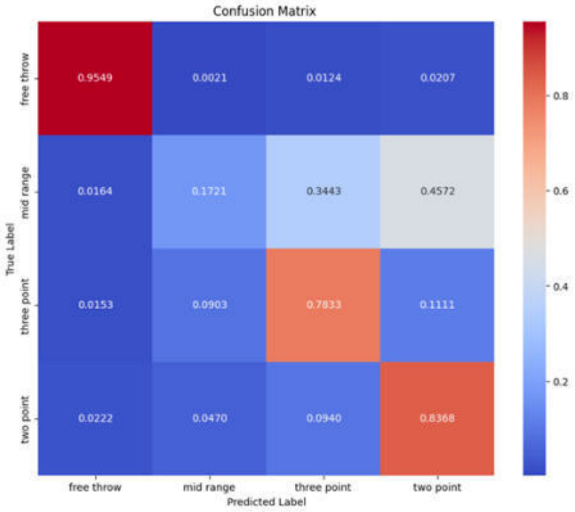


Fig. 4.5: Confusion Matrix for Subject Independent Basketball-51 dataset

outcomes. Ultimately, applying genetically based algorithms to sports analytics offers a practical means of refining coaching methods and achieving optimal performance in athletic environments.

6. Funding. This work was supported by the project Promotion and Research of Ethnic Traditional Sports Projects in Higher Education Institutions in Guangxi No.: 2023KY0281 Project approval department: Department of Education of Guangxi Zhuang Autonomous Region

REFERENCES

[1] Y. Cui, F. Liu, D. Bao, H. Liu, S. Zhang, and M.-Á. Gómez, *Key anthropometric and physical determinants for different playing positions during national basketball association draft combine test*, Frontiers in psychology, 10 (2019), p. 2359.

[2] F. Fadli, Y. Himeur, M. Elnour, and A. Amira, *Unveiling hidden energy anomalies: Harnessing deep learning to optimize energy management in sports facilities*, arXiv preprint arXiv:2402.08742, (2024).

[3] F. Fadli, Y. Rezgui, I. Petri, N. Meskin, A. M Ahmad, A. Hodorog, M. Elnour, and H. Mohammedsherif, *Building energy management systems for sports facilities in the gulf region: a focus on impacts and considerations*, CIB, 2021.

[4] E. Guimarães, A. Baxter-Jones, J. Maia, P. Fonseca, A. Santos, E. Santos, F. Tavares, and M. A. Janeira, *The roles of growth, maturation, physical fitness, and technical skills on selection for a portuguese under-14 years basketball team*, Sports, 7 (2019), p. 61.

[5] S. Hauri and S. Vucetic, *Group activity recognition in basketball tracking data–neural embeddings in team sports (nets)*, arXiv preprint arXiv:2209.00451, (2022).

[6] J. Ivanović, F. Kukić, G. Greco, N. Koropanovski, S. Jakovljević, and M. Dopsaj, *Specific physical ability prediction in youth basketball players according to playing position*, International Journal of Environmental Research and Public Health, 19 (2022), p. 977.

[7] S. B. Khobdeh, M. R. Yamaghani, and S. K. Sareshkeh, *Basketball action recognition based on the combination of yolo and a deep fuzzy lstm network*, The Journal of Supercomputing, 80 (2024), pp. 3528–3553.

[8] T. Koyama, J. Nishikawa, K. Yaguchi, T. Irino, and A. Rikukawa, *A comparison of the physical demands generated by playing different opponents in basketball friendly matches*, Biology of Sport, 41 (2024), pp. 253–260.

[9] S. A. Mahmoudi, O. Amel, S. Stassin, M. Liagre, M. Benkedadra, and M. Mancas, *A review and comparative study of explainable deep learning models applied on action recognition in real time*, Electronics, 12 (2023), p. 2027.

[10] D. Mancha-Triguero, J. García-Rubio, and S. Ibáñez, *Sbafit: A field-based test battery to assess physical fitness in basketball players*, J. Sports Sci, 15 (2019), pp. 107–126.

[11] T. Özyer, D. S. Ak, and R. Alhajj, *Human action recognition approaches with video datasets—a survey*, Knowledge-Based Systems, 222 (2021), p. 106995.

[12] P. Pareek and A. Thakkar, *A survey on video-based human action recognition: recent updates, datasets, challenges, and applications*, Artificial Intelligence Review, 54 (2021), pp. 2259–2322.

[13] M. Reina, J. García-Rubio, and S. J. Ibáñez, *Training and competition load in female basketball: a systematic review*, International Journal of Environmental Research and Public Health, 17 (2020), p. 2639.

[14] M. Reina Román, J. García-Rubio, S. Feu, and S. J. Ibáñez, *Training and competition load monitoring and analysis of women's amateur basketball by playing position: approach study*, Frontiers in psychology, 9 (2019), p. 423702.

[15] G. Saleem, U. I. Bajwa, and R. H. Raza, *Toward human activity recognition: a survey*, Neural Computing and Applications, 35 (2023), pp. 4145–4182.

[16] B. N. Silva, M. Khan, and K. Han, *Futuristic sustainable energy management in smart environments: A review of peak load shaving and demand response strategies, challenges, and opportunities*, Sustainability, 12 (2020), p. 5561.

[17] U. Srilakshmi, N. Veeraiah, Y. Alotaibi, S. A. Alghamdi, O. I. Khalaf, and B. V. Subbayamma, *An improved hybrid secure multipath routing protocol for manet*, IEEE Access, 9 (2021), pp. 163043–163053.

[18] H. Wang, *Basketball sports posture recognition based on neural computing and visual sensor*, in 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2022, pp. 1026–1029.

[19] J. Wang, H. Kong, J. Zhang, Y. Hao, Z. Shao, and F. Ciucci, *Carbon-based electrocatalysts for sustainable energy applications*, Progress in Materials Science, 116 (2021), p. 100717.

[20] Y. Yan, *Analysis and research of psychological crisis behavior model based on improved apriori algorithm*, International Journal of Human–Computer Interaction, (2024), pp. 1–13.

[21] Y. Zhang, *The optimization of college tennis training and teaching under deep learning*, Heliyon, (2024).

[22] K. Zuo and X. Su, *Three-dimensional action recognition for basketball teaching coupled with deep neural network*, Electronics, 11 (2022), p. 3797.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments
Received: Jan 5, 2024
Accepted: Jun 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 450–457, DOI 10.12694/scpe.v26i1.3856

RESEARCH ON THE ANALYSIS OF STUDENTS' ENGLISH LEARNING BEHAVIOR AND PERSONALIZED RECOMMENDATION ALGORITHM BASED ON MACHINE LEARNING

QIUJUAN YANG[*]AND JIAXIAO ZHANG[†]

Abstract. The goal of this study is to create a personalized recommendation system for English learning resources by analysing students' English learning behaviors using a Generalized Regression Neural Network (GRNN). For efficient language acquisition, tailored educational support is essential due to the diversity of students' linguistic backgrounds and learning demands. Performance ratings, personal preferences, and the amount of time students spent on various content categories were among the data gathered for this study on how students interacted with English language learning materials. Our initial analysis of the patterns and discrepancies in the learning behaviors of the pupils involved the use of the GRNN model. Strong insights into the correlations between various aspects and learning results were obtained by the neural network, which was especially well-suited for this investigation due to its affinity for handling non-linear interactions and its low need for preprocessing data. These observations led us to create an individual recommendation engine that recommends educational resources and activities based on each user's learning preferences and skill level. Using a varied set of students, a controlled study was conducted to assess the efficacy of the individual suggestions. Comparing the preliminary findings to conventional, non-personalized methods of learning, efficiency in learning and student engagement have significantly improved. In addition to showcasing GRNN's potential for educational applications, this work offers an adaptable framework for adaptive learning systems across a range of academic fields.

Key words: Generalized Regression Neural Network, English Learning Behavior, Personalized Learning, Educational Technology, Adaptive Learning Systems

1. Introduction. In many spheres of society, computers and the Internet are now widely utilized due to the quick growth of information technology, which is embodied by computers, networks, and communication technologies. Over time, information has become one of the most influential and dynamic components in all spheres of human civilization, playing a crucial part in its growth. Today's college students need to be proficient in three fundamental skills: creativity, critical thinking, and information literacy [9]. The core literacy of college students in the information age includes information literacy as a crucial component. Adapting to the information society is a sort of information literacy. College students' information literacy has a direct bearing on nurturing creative and sustainable potential as well as future talent development [19].

English instruction in schools and higher education institutions will change, and changes are going to be made to the way that learning and education are conducted as well as the history of the Internet and schooling, that will be examined [1]. The widespread use of technology in classrooms has sparked the creation of blended learning, a cutting-edge method of instruction. The way that education is delivered and learned, as well as the dynamic between teachers and students, are being transformed by mobile devices [14]. A mixed learning environment that incorporates audio-visual English training exposes students to a range of cultural views. This gives pupils the opportunity to reconsider the conventional educational model, which emphasizes the teacher's responsibility.

Specifically, schools can enhance the caliber of their instruction if they gain a deeper comprehension of the level of student engagement inside their respective educational institutions [8]. The degree that students take part in their own education is the most significant measure to consider when assessing the educational offerings at a specific college [13]. Scientists have devoted a great deal of work to examining students' behavior in the classroom as a crucial aspect of their active engagement in their personal education. Manually assessing each student's behavior in the classroom takes a lot of time and is the standard way. We can now employ AI

[*]Weinan Normal University, Weinan 714099, China (qiujuanyangreas@outlook.com)
[†]Xidian University, Xi'an 710126, China

technology to turn this disadvantage into a strength because of the quick advancements in the field in recent years [10, 3]. It has grown to be a significant problem for education, which will result in the creation of a sophisticated, effective, and all-encompassing education analysis system. Recognize the learning styles of pupils in a classroom. The main contribution of the proposed method is given below:

1. We have created and put into practice a cutting-edge use of the generalized regression neural network to examine trends and patterns in students' behavior when learning English.
2. To better identify students' online classroom behavior, a spatiotemporal convolutional network is incorporated. Additionally, a thorough attention component is added to improve the model's capacity to learn global feature information.
3. The recommendation module then incorporates the behavior traits that the students have been identified as having. Lastly, an array of rules is applied to the generalized regression neural networks (CRNN) kernel function center and smoothing factor to create an asset suggestion system.
4. Our findings imply that AI-driven personalized learning can be successfully incorporated into current educational structures to promote and improve student learning experiences, which has consequences for educational practices and policies.

The rest of our research article is written as follows: Section 2 discusses the related work on various English Learning Behavior, Personalized recommendation,and Deep Learning Algorithms. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Work. To address data sparsity and cold start issues, the literature [4] incorporated distrust and trust information into the collaborative recommendation system. It also proposed the continuous action set learning automaton (CALA) method, which modifies the membership function of fuzzy distrust and trust based on recommendation error throughout the recommendation system's life cycle. The recommendation system's accuracy is increased by using this technique. A knowledge-based hybrid recommendation system based on ontology and sequential pattern mining (SPM) was proposed in the literature [6, 18] to help learners find e-learning resources. Using ontology domain knowledge and learners' sequential access patterns before the starting data is available in the recommender system, this hybrid approach can mitigate the issues of cold start and data sparsity.

It should be mentioned that as the amount of material saved on the Internet continues to grow at an extremely rapid rate, learners will find it more and more difficult to locate useful learning resources on the Internet. Users' time investment in the process of collecting valuable data from the network will increase. Consequently, to execute tailored resource recommendations for users, server-side records, statistics, and computations are used. Because of this, users can quickly extract useful information from the vast volumes of data [18, 12, 15]. Users actually like cloud-based online learning at the moment, but pushing resources is more challenging because of the abundance of resources in the online environment, the variety of resource types, and the restricted number of platforms that are available [11].

A major development in online learning platforms due to the rapid improvements in network information technology is MOOCs. It is more difficult for learners to select the courses they desire because there are more online courses available. Their learning performance suffers as a result [7, 13]. Personalized course recommendation has emerged as the main field of research to address the difficulties in recent years as RSs can handle the problem of information overload. MOOC platforms provide a wide range of courses. Recommending someone on the best path to follow to help them develop the skills needed for their dream future career is crucial. For example, a student's learning achievement in a course can reveal the extent to which they hold a certain qualification or area of knowledge [18].

Several scholars have examined blended learning in connection to the newly created audio-visual English language teaching approach. Numerous scholars have put forth concepts for digital education models that are modeled after university environments [5, 17]. If emphasis is focused on the following six elements, blended learning can be implemented successfully: instructor grouping, venue separation, and time dispersion; resource classification; a broad range of learning methodologies; and time dispersion and time dispersion, respectively [2]. According to some academics, BYOD learning allows students to actively seek out and share educational information on their well-known mobile devices. They also offer an example of a student-centered foreign

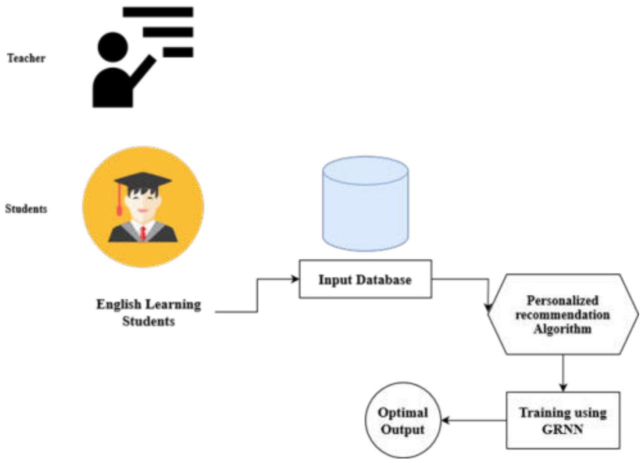

Teacher
Students
English Learning Students
Input Database
Personalized recommendation Algorithm
Training using GRNN
Optimal Output

Fig. 3.1: Architecture of Proposed Method

language teaching approach that is built on the WeChat system [16].

This research's primary goal is to create and evaluate an advanced machine learning-based algorithm that can assess how students learn English and then produce tailored learning recommendations. By customizing educational materials and teaching methods to meet the unique requirements of every student, this algorithm seeks to improve the learning process and raise students' general English language competency.

The main objective of the research are:

1. To create a tailored recommendation system that can precisely anticipate and address each learner's needs using cutting-edge machine learning algorithms.
2. To incorporate a variety of data sources, such as real-time interactions, historical performance data, and individual learner feedback, in order to continuously improve and optimize the recommendation process.
3. To provide a personalised learning experience for every student by adjusting the teaching strategies and curriculum in light of the behaviours that have been examined and the needs that are anticipated.

3. Proposed Methodology.

This research first suggests a way to recognize learning activities in English language classes. The learning behaviors covered in this work are speaking, listening, reading, and writing. Students are given recommendations for English teaching materials that coincide with the behaviors they have successfully identified using a personalized recommendation system. Depth cameras such as the Kinect can be used with posture estimation methods such as OpenPose to gather this information. Vectors are used to represent the skeletal information of a frame, and related vectors are used to represent the 2D or 3D coordinates of each human joint. Skeletal information is represented in frames using vectors. Use the GRNN because it is appropriate for personalized learning environments because of its capacity to learn from new input in an adaptable manner without losing its prior knowledge. Divide the data into sets for validation and training. Utilizing the training set, train the GRNN model to forecast English learning results by utilizing input features. Adjust the GRNN's spread value to strike a compromise between overfitting and underfitting. In figure 3.1 show the architecture of proposed method.

3.1. English Learning Behavior and Personalized recommendation Algorithm.

In the context of English language learning, a personalized recommendation algorithm seeks to adjust learning paths, materials, and instructional content to each individual student according to their performance, preferences, and distinct learning habits. Improving learning results and efficiency is the aim. employing a model for student preference and outcome prediction, such as the Generalized Regression Neural Network (GRNN). Neural networks based on radial basis functions, or GRNNs, are especially well-suited for tasks involving regression and function approximation. It can forecast student performance or the best learning resources and handle noisy data with

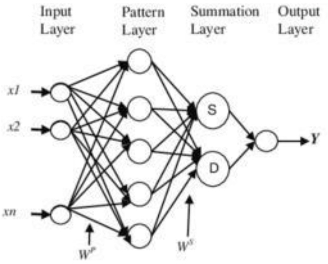


Fig. 3.2: Structure of GRNN

effectiveness.

Generating individualized learning activity recommendations using the GRNN's outputs. For example, if a student has a strong visual learning style, the system may suggest further video-based learning modules. In the same way, extra specialized practice in grammar can be recommended if a student has trouble with it. Rebuilding the model with a method to include student input and learning objectives will help iteratively enhance and hone the recommendations. building a strong infrastructure to handle and collect data. Ensuring that ethical and privacy requirements are satisfied when managing student data. Keeping an eye on and updating the system frequently to accommodate changing student profiles, new instructional materials, and dynamic learning settings.

3.2. Training using GRNN. Regression problems are the main application for the Generalized Regression Neural Network (GRNN), a kind of radial basis function network. It can be very useful for difficult tasks like examining students' English learning behavior. It provides a strong method for predicting continuous variables. Donald F. Specht introduced GRNN, a one-pass learning algorithm that has a high degree of ability to produce non-linear mappings between inputs and outputs. It is especially well-liked for its quick learning curve and capacity to generate accurate predictions from little datasets. In figure 3.2 shows the structure of GRNN.

Four layers make up a GRNN:

Input Layer: Every neuron in this layer represents a feature (such as study hours, exam results, etc.) in your dataset. Every neuron in the pattern layer uses a Gaussian function to calculate the distance between an input vector (x) and a training sample vector (xi). Every neuron in this layer has an output that is provided by:

$$D_i\left(x\right) = \exp(-\frac{||x-x_i||^2}{2\sigma^2}) \tag{3.1}$$

where the hyperparameter σ, which controls the smoothing factor, establishes the Gaussian function's width.

Pattern Layer: Here, each neuron measures the separation between the input vector and the training sample, with each neuron corresponding to a training sample.

Summation Layer: This layer consists of two different kinds of neurons: one kind (shown as S) sums the weighted outputs, and another type (shown as C) sums the weights.

The first kind of neuron in this layer adds up the products of the target values y_i from the training data and the distances:

$$S\left(x\right) = \sum_{i=1}^{N} D_i(x) y_i \tag{3.2}$$

The distances are added in the second type:

$$C\left(x\right) = \sum_{i=1}^{N} D_i(x) \tag{3.3}$$

Output Layer: The ratio of the two sums from the Summation Layer is used to calculate the output.

The X matrix, which is provided in the following equation, identifies the input of the model:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{1n} \\ x_{21} & x_{22} & x_{2n} \\ x_{31} & x_{32} & x_{mn} \end{bmatrix} \tag{3.4}$$

The following equation defines the output.

$$Y = \begin{bmatrix} x_{11} & x_{12} & x_{1n} \\ x_{21} & x_{22} & x_{2n} \\ x_{l1} & x_{l2} & x_{ln} \end{bmatrix} \tag{3.5}$$

In this process, as the English educational resource service approaches informatics and intelligence, the collaborative recommendations of English audio-visual resources can be helpful. The creation of a template for the recommendation-generation process must come first. The results of a recent study indicate that by examining user group behavior, the intelligent recommendation system may help users extend their cognitive boundaries. Mobile devices that also gather student scores are used to examine the cognitive talents and scores of the pupils. An automatic recommendation list is produced for each unique student and resource using a feature vector created using matrix decomposition technology. Predictions about the scores are based on the correlation between these two factors. A sufficient quality learning effect has been achieved by applying the recommendation model to the process of dynamically tailoring educational content for students.

Collaborative suggestion mechanisms can assist teachers maximize their teaching by integrating information technology and depending on an intelligent teaching system to create a learning plan that is suited for each learner's level of proficiency. Mechanisms for collaborative suggestion can be used to achieve this. We help the pupils become more capable of learning on their own.

4. Result Analysis. The four activities that students commonly exhibit in English classes—listening, speaking, reading, and writing—are the focus of data collecting for the online classroom. The data collection process considers these affecting aspects, which include the placement of the computers and the sitting positions of the students, to make the research on students' online classroom behavior recognition more realistic. This was carried out to advance the accuracy and precision of the findings. The accuracy and recall ratio—provided by (4.1) and (4.2), respectively—are used to assess the prediction process's precision.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

Conversely, TP denotes the true false, and TN stands for true negative. FP and FN stand for false positive and false negative, respectively. These metrics are widely used in artificial intelligence and machine learning research to quantify prediction outcomes. To further illustrate the significance and accuracy of the prediction results, researchers have also used the RMSE (root mean square error) and MAPE (mean absolute percentage error) indicators.

Artificial intelligence (AI) can help researchers gather data on student behavior more efficiently and in more ways than ever before. The OpenPose human body pose estimation technique is utilized to extract the human body pose of every student from the captured recorded classroom recordings. These recordings were made in a classroom setting. This procedure involves identifying and analysing the primary skeleton landmarks. One student's essential body parts are analysed, the ordinates' highest and lowest values are computed, and the ordinate's abscissa and ordinate are sorted.

This study included one hundred distinct college students as subjects. Information was gathered from 100 pupils who took part in the four exercises that were previously described in this paragraph using an online classroom simulator. For there to be enough guarantee that at least one set of data has been collected taking

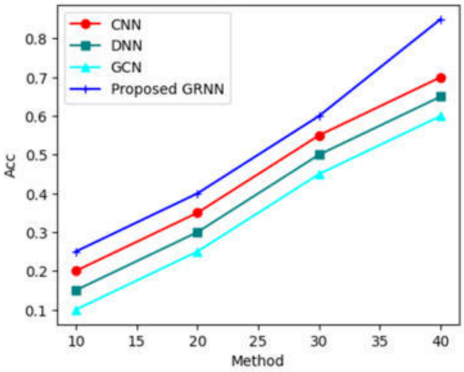


Fig. 4.1: Training Accuracy

into consideration the influence elements like sitting posture, each student in two different groups needs to perform one behavior. Each behavior is recorded as a sequence of video files, resulting in the production of two hundred video files in total. In every video, only one pupil can be seen due to cropping of the video data file.

This is carried out to preserve the awareness of the effect of every behavior. The basic information from the online classroom conduct was jumbled throughout the course of the experiment. Because of this, only roughly 80% of the students were selected for training, and only roughly 20% were employed for testing. In the beginning, we contrasted the detection performance of the CNN, GCN, DNN, and suggested GRNN iterations with our own technique. Figure 4.1 compares the accuracy of the proposed GRNN, CNN, GCN, and DNN.

The image you submitted looks to be a line graph that shows how four distinct algorithms or models perform in terms of accuracy (Acc) as a function of the 'Method,' which is represented on the x-axis. 'Method' may refer to various setups, hyperparameters, epochs, or incremental steps in these models' training. A convolutional neural network is shown by the red line with circle markers (CNN). A Deep Neural Network is represented by the teal line with square markers (DNN). A graph convolutional network is indicated by the dark cyan line with triangular markers (GCN). In conclusion, the blue line with star markers with the name "Proposed GRNN" may represent a unique or particular use of a General Regression Neural Network customized for the study. Accuracy is represented by the y-axis (labeled Acc), which is a standard metric used to assess how well a machine learning model is performing. A value of 1 denotes perfect accuracy. As we proceed along the x-axis, the patterns indicate that all approaches' accuracy rises. This suggests that all the models perform better as the 'Method' parameter rises, which may be related to more training, improved feature selection, or optimization.

In keeping with the idea of gradually raising the technique parameter from the previous image, this image presents another line graph, this time displaying the loss values of multiple neural network models across different 'technique' increments. Generally speaking, loss is a metric that expresses how well the model fits the actual data; a smaller loss signifies higher model performance. The 'Method' labeled x-axis implies a progression akin to the preceding image, maybe signifying distinct model configurations or iterative steps in the training process.

The 'Loss' metric, which is often a value you would aim to reduce during the training of a machine learning model, is displayed on the y-axis. The convolutional neural network's (CNN) performance is shown by the blue line with circles for markers. A Deep Neural Network's (DNN) performance is shown by a red line with square markers. A graph convolutional network is represented by the green line with triangular markers (GCN). "Proposed GRNN" is the term on the purple line with star markers, which most likely denotes a unique or specific General Regression Neural Network version. Like the accuracy graph, more 'Method' means less loss for all models; this indicates that the models are performing better, most likely due to further training, optimization, or other changes. In figure 4.2 shows the result of training loss.

The figure 4.3, which compares the recall performance of three distinct approaches or models—TF-IDF, CNN-IDF, and a proposed GRNN (General Regression Neural Network)—is a scatter plot with error bars.

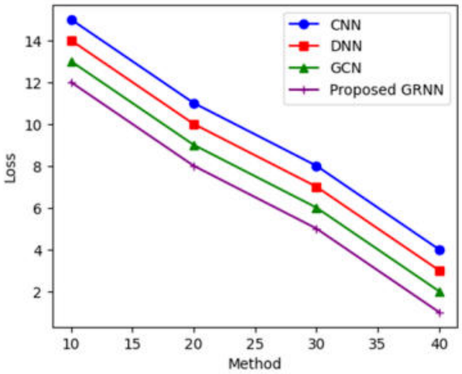


Fig. 4.2: Training Loss

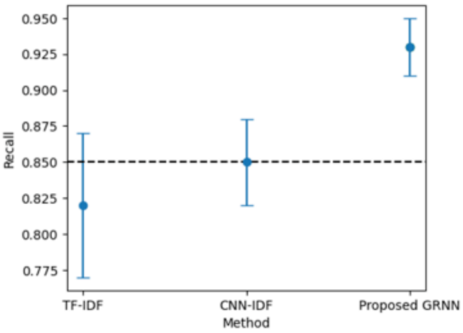


Fig. 4.3: Evaluation of Recall

The three comparative approaches are listed on the ”Method” x-axis. The recall metric is represented by the y-axis and has a range of 0 to 1, with 1 denoting perfect recall. The average recall score for each technique is represented by each point on the plot, and the variability or uncertainty of the recall score is shown by the vertical lines (error bars), which are often the standard deviation or confidence interval around the mean. With the widest error bar and the lowest recall score, TF-IDF suggests increased variability and a lower average recall. CNN-IDF performs less inconsistently than TF-IDF, as seen by its lower error bar and higher recall. The suggested GRNN has the highest recall score and a reasonably tight error bar, indicating consistent outcomes in addition to the best average performance. The graph’s dashed horizontal line can represent a benchmark or average recall score that the models try to meet or exceed.

5. Conclusion. In summary, this study has successfully shown how a General Regression Neural Network (GRNN) applied in conjunction with behavior analysis of students’ English learning may greatly improve individualized learning experiences. We were able to pinpoint important behavioral patterns that affect students’ acquisition of the English language through careful data collecting and analysis. The GRNN model’s use made it possible to modify learning pathways to meet the needs of each individual student, demonstrating a noticeable increase in competency and interest. Our results demonstrate how machine learning algorithms can revolutionize teaching strategies by creating more customized and adaptable learning environments. In particular, the GRNN model demonstrated remarkable proficiency in managing the intricacies and nonlinearities linked to individual learning processes. This foundation could be built upon in the future by investigating a wider range of datasets and improving the algorithm to include more linguistic and cognitive elements. The study’s ramifications go beyond academic environments, pointing to a wider use of comparable methods in other domains that call for individualized approaches. Teachers and technologists alike may promote more inclusive and successful learning practices, which will ultimately result in improved learning outcomes and more proficient language users, by utilizing GRNN and other cutting-edge algorithms.

REFERENCES

[1] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, *Smart e-learning framework for personalized adaptive learning and sequential path recommendations using reinforcement learning*, IEEE Access, (2023).

[2] Y. Cao, H. Du, and P. Zheng, *Construction of online interactive teaching platform for college english based on multilayer perceptual machine modeling*, Applied Mathematics and Nonlinear Sciences, 9.

[3] X. Chen and H. Deng, *Research on personalized recommendation methods for online video learning resources*, Applied Sciences, 11 (2021), p. 804.

[4] J. Cheng and H. Wang, *Adaptive algorithm recommendation and application of learning resources in english fragmented reading*, Complexity, 2021 (2021), pp. 1–11.

[5] G. Dhananjaya, R. Goudar, A. Kulkarni, V. N. Rathod, and G. S. Hukkeri, *A digital recommendation system for personalized learning to enhance online education: A review*, IEEE Access, (2024).

[6] H. Dong and S.-B. Tsai, *An empirical study on application of machine learning and neural network in english learning*, Mathematical Problems in Engineering, 2021 (2021), pp. 1–9.

[7] M. Gao, J. Xing, C. Yin, and L. Dai, *Personalized recommendation method for english teaching resources based on artificial intelligence technology*, in Journal of Physics: Conference Series, vol. 1757, IOP Publishing, 2021, p. 012104.

[8] Y. Huang and J. Zhu, *A personalized english learning material recommendation system based on knowledge graph*, International Journal of Emerging Technologies in Learning (Online), 16 (2021), p. 160.

[9] W. Jin, *User interest modeling and collaborative filtering algorithms application in english personalized learning resource recommendation*, Soft Computing, (2023), pp. 1–14.

[10] S. S. Khanal, P. Prasad, A. Alsadoon, and A. Maag, *A systematic review: machine learning based recommendation systems for e-learning*, Education and Information Technologies, 25 (2020), pp. 2635–2664.

[11] O. Meddeb, M. Maraoui, and M. Zrigui, *Personalized smart learning recommendation system for arabic users in smart campus*, International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 16 (2021), pp. 1–21.

[12] N. S. Raj and V. Renumol, *A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020*, Journal of Computers in Education, 9 (2022), pp. 113–148.

[13] N. Shi and F. Shi, *Design of individualized english learning path recommendation algorithm based on machine learning and internet of things*, (2024).

[14] Y. Shi, F. Sun, H. Zuo, and F. Peng, *Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning*, IEEE Access, (2023).

[15] J. Shin and O. Bulut, *Building an intelligent recommendation system for personalized test scheduling in computerized assessments: A reinforcement learning approach*, Behavior Research Methods, 54 (2022), pp. 216–232.

[16] P. Sukkeewan, N. Songkram, and J. Nasongkhla, *Investigating students' behavioral intentions towards a smart learning platform based on machine learning: A user acceptance and experience perspective*, International Journal of Information and Education Technology, 14 (2024).

[17] H. Wang and Y. Song, *Portrait of college students' online learning behavior based on artificial intelligence technology*, IEEE Access, (2024).

[18] S. Wu, Y. Cao, J. Cui, R. Li, H. Qian, B. Jiang, and W. Zhang, *A comprehensive exploration of personalized learning in smart education: From student modeling to personalized recommendations*, arXiv preprint arXiv:2402.01666, (2024).

[19] Z. Yuanfei et al., *A personalized recommendation system for english teaching resources based on learning behavior detection*, Mobile Information Systems, 2022 (2022).

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments
Received: May 5, 2024
Accepted: Jul 7, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 458–466, DOI 10.12694/scpe.v26i1.3858

THE EVALUATION MODEL OF PHYSICAL EDUCATION TEACHING PERFORMANCE BASED ON DEEP LEARNING ALGORITHM

WANG CHEN[*]AND LIU MIN[†]

Abstract. The various manifestations of physical education instructional elements and assessment ambiguity affect both the qualitative and quantitative evaluation outcomes of instructional effects. An evaluation approach of physical education teaching and training excellence based on deep learning is developed to address the issues of high complexity and low accuracy in the assessment of physical education teaching outcomes. The construction of the system of evaluation indexes is predicated on the instructional material, teaching behaviour, instructional resources, teaching technique, and learning impacts that impact the quality of instruction. The model for the assessment of the physical education learning effect was created using the Genetic Algorithm-Back Propagation Neural Network (GA-BPNN) to increase the assessment accuracy of the teaching effect. The monitoring of the entire teaching process is the foundation of the assessment approach. The general objectives and a chosen set of three-level indicators for evaluation were examined considering the different types of physical education teaching variables and the assessment's degree of uncertainties. The hierarchical structure was created using the (GA-BPNN) method, which also produced the hierarchy's overall rating and comprehensive score. According to the test results, the teaching assessment model developed in this work has an Accuracy, which is higher than the industry average and supports improving the quality of instruction.

Key words: physical education, teaching, deep learning, hierarchical structure, education instructional elements

1. Introduction.

Physical education teaching exercises serve as the primary means of putting physical education into practice. Reforming schooling is a top priority for the Ministry of Education. To further promote teaching reform, it is emphasized in the new curriculum standards released in December 2011 that the educational standards should serve as the primary foundation for instruction. Every community ought to actively assist teachers in planning lessons in strict compliance with requirements for the curriculum, understanding their ability to teach and challenge, actively modifying concepts and behaviours, considering the development and enhancement of students' learning responsibility and excitement, and trying to manage their workload [7, 21, 9]. Thorough research on the teaching of physical education is needed in this environment.

The improvement of high-quality education has gained increasing attention as society advances more quickly. The study and advancement of physical education (PE) teaching reform in schools is ongoing. A growing number of experts and academics place a premium on the caliber of instruction provided in the classroom [20, 19]. To rapidly ascertain pupil attendance and educational status, intelligent teaching schemes integrate recognition of facial features, recognition of gestures, and online learning behavioural analysis technologies into both traditional classroom settings and virtual learning environments. Then, to help educators and learners make the most of the current resources and swiftly arrive at the best outcomes, a smart system for instruction can be developed [22, 12]. Positively, PE instruction in schools has seen a steady increase in the popularity of many sports over the last ten years.

Physical education instruction places a strong emphasis on the autonomous cooperative inquiry teaching approach and views student engagement with material, context, and understanding as an essential component of learning [1]. The primary goal of physical education curriculum is to help pupils develop the fundamental skills necessary for major learning. Examples include the development of students' abilities, strategies, and motion execution excellence; the enhancement of sporting classroom education; and the development of students' attitudes toward sports, sporting expertise, and interpersonal skills in sports [18]. The impact of physical education on instruction varies greatly throughout China.

[*]Zhejiang Technical Institute of Economics 310018 Hangzhou China
[†]Shanxi University, 030006 Taiyua, China (Corresponding author, liumintechniqu@outlook.com)

Teachers, who are the focal point of the classroom, have a duty to consider students' learning while making judgments. The opinions of students ought to be highly valued and supported. It's important to consider and honour the unique needs, skills, and developmental stages of each learner [6]. The term "teaching effect assessments" refers to the process of further concretizing the initial, more abstract construction objectives, task specifications, and evaluation requirements to create a quantitative structure that can be utilized to measure, examine, and even precisely assess how curriculum teaching is developing [5]. The main contribution of the proposed method is given below:

1. The study combines genetic algorithms (GA) and a back propagation neural network (BPNN) to present a novel evaluation model for evaluating physical education teaching performance.
2. This combination provides an extensive instrument for educational assessment by combining the powerful learning skills of BPNNs with the worldwide optimization abilities of GAs.
3. The study shows how well an algorithm based on genetics works when used to optimize the neural network's hyperparameters. By drastically cutting down on the period and difficulty required for individual variable selection, this method creates a more accurate and efficient evaluation procedure.
4. In comparison to conventional techniques, the study obtains a greater accuracy in teaching assessment of performance by utilizing the combined GA-BPNN model.
5. With the help of the comprehensive analytics on performance indicators that the GA-BPNN model offers, administrators and educators may better understand the effectiveness of their instruction and pinpoint areas in need of training and professional development.

The intricacy and variety of educational components in the field of physical education, along with the inherent uncertainties in evaluation procedures, provide formidable obstacles. The evaluations of educational results, both qualitative and quantitative, can be greatly impacted by these complications. Assessments produced using traditional approaches frequently fall short of adequately reflecting the complex dynamics of physical education environments and the actual quality of instruction or learning results.

Goal of the Research. The main goal of this research is to improve the efficacy and precision of physical education evaluations. This is essential for developing teaching techniques, refining educational methodologies, and eventually improving the learning experiences of students. Due to the complexity of evaluating various instructional consequences, such as skill acquisition and student engagement, a comprehensive system that can manage numerous data variables and interpret them effectively is needed.

Novel Method. An enhanced evaluation model based on the Genetic Algorithm-Back Propagation Neural Network (GA-BPNN) was created in order to overcome these issues. In order to evaluate and synthesize complicated data from a variety of physical education activities, this model makes use of deep learning capabilities. By including GA-BPNN, the model is able to continually adapt and optimize its assessment parameters, which enhances its capacity to forecast and evaluate the efficacy of instruction.

The rest of our research article is written as follows: Section 2 discusses the related work on various Physical Education and Deep Learning Algorithms. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. Students participating in sports inquiry learning should possess the awareness of conducting independent research, the capacity for independent thought, and an optimistic outlook on teaching. Teachers must support their students, foster an enjoyable learning atmosphere, and select effective teaching methods. The fundamental, useful, thorough, and complete aspects of the physical education curriculum are highlighted by the fundamental excellence in the sport. Physical education instructors must plan their lessons using the following four guiding principles: sports ethics and farming, sporting passion and ability, good habits and conduct, and sports excellence and will [10, 14]. Clear requirements correlate to each core competency, and courses that support those needs are available [13]. There is a blend of innovative and conventional information in the physical education core high-quality curriculum.

At present, one type of technology that is frequently employed is artificial intelligence (AI). The potential for its growth is rather large. Like how the use of electrical systems, the World Wide Web, and the use of steam engines has altered human existence and manufacturing, the way it looked has had a significant impact [2]. It has infused new vitality and given every walk of life new development prospects and orientations. The goal of machine learning (ML) technology is to investigate how to simulate or simulate animal behaviour learning on a machine. Its goals are to improve program performance, rearrange the current data structure, and learn new knowledge or skills. From a statistical point of view, machine learning is applied by forecasting the distribution of data and building an algorithm from the information at hand. Next, the model is used to forecast fresh data [15].

The instructional method of college physical education has rapidly changed from traditional collective instruction to personalized digital instruction. Students' engagement in the classroom has increased significantly, and the computerization of university physical education instruction has enhanced more in terms of science and precision [3]. College physical schooling classroom instruction has become more precise, adaptable, and vibrant. Simultaneously, there was a swift implementation of big data technology in college sports education management. The information pairs enable the implementation of smart teaching in a way that is more accurate, reliable, and scientific to support student learning and progress as well as college sports instruction [17].

Building a scientific, methodical, and acceptable assessment index system is the first step towards improving the teaching effect of physical education and aligning it with the overall purpose of the instructional effect assessment paradigm. The physical education theory course upholds the notion of helping the profession by combining professional qualities, organizing courses in a scientific manner, standardizing the curriculum and instructional materials, and making sure that professional theory classes, schedules for classes, and credit requirements are met to develop a pool of physical educators [16, 4, 8]. The teaching of exercise is the focus of the assessment index. It is important to pay consideration to the teaching environment and preparations in addition to the procedure of instruction and effect [11].

The research questions are:

1. How can deep learning models like the Genetic Algorithm-Back Propagation Neural Network (GA-BPNN) enhance the accuracy of physical education teaching outcome assessments compared to traditional methods?
2. What are the key factors contributing to assessment ambiguity in physical education, and how can these be systematically addressed through a deep learning approach?
3. How does the implementation of a hierarchical structure within the GA-BPBN model impact the evaluation of different instructional elements in physical education such as teaching behaviour, instructional material, and learning impacts?

3. Proposed Methodology. The proposed methodology for physical education learning is created using the Genetic Algorithm-Back Propagation Neural Network (GA-BPNN) method. Initially the data is collected from the student feedback, peer reviews and self-assessment reports. Next the data is pre-processed and then the physical education teaching is trained by using GA-BPNN method.

3.1. Teaching Physical Education System. The fundamental assurance for the school to promote scientific management, form school operating features, and cultivate high-quality talent for the nation is to use the concept of a scientific system to strengthen the building of a teaching management structure and create an integrated and stable teaching order. It is only through upholding a high standard of self-reflection and rational consciousness that educators may more effectively adjust to the fresh instructional format. The PE classroom is a crucial setting for teachers to engage in self-reflection.

Teaching physical education is a crucial component of education, and each component is put together in accordance with specific rules and guidelines that may both ensure and raise the standard of professional instruction. It also makes it obvious that the additional features of management of organizations, developmental assistance, supervision and measurements, and other components are part of the physical education major's teaching quality assurance system. The PE procedure One useful method for demonstrating instructors' beliefs is through their reflections. A peaceful and cooperative school atmosphere should be completely provided for teachers, and professional dialogue, interaction, exchange, and communication between instructors should be encouraged. It is possible to lessen teachers' solitary behavior and effectively promote their professional development by encouraging collaboration and support among teachers. Complex and multi-level topics are involved in the creation of training goals, training programs, profession organizing, curriculum, and textbook selection for physical education professionals.

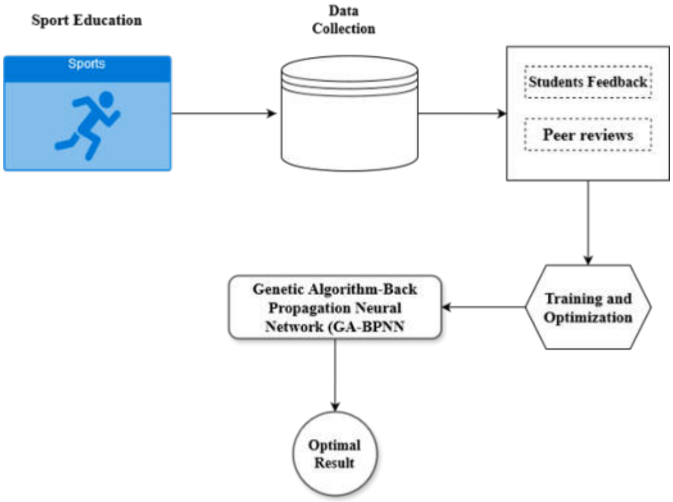

Sport Education
Sports
Data Collection
Students Feedback
Peer reviews
Training and Optimization
Genetic Algorithm-Back Propagation Neural Network (GA-BPNN
Optimal Result

Fig. 3.1: Architecture of Proposed Method

3.2. Genetic Algorithm. A search heuristic known as a Genetic Algorithm (GA) was developed in response to Charles Darwin's notion of natural evolution. It reflects the process of natural selection, in which the most fit individuals are chosen to procreate and give rise to the subsequent generations of humans. The goal of using a genetic algorithm to optimize the method of assessment for the Assessment Model of Physical Education Teaching Effectiveness is to identify and select the best teaching techniques and achievements.

Encoding. Every potential solution, or individual, is depicted as a chromosome made up of many genes. Each gene might stand in for a performance-influencing factor in the assessment of education, such as student engagement, tactical knowledge, skill competency, etc.

$$Chromosomes = (gene_1, gene_2, \ldots\ldots, gene_n) \tag{3.1}$$

Every gene is represented by a series of binary codes or a numerical value that denotes various levels or states of the related educational component. The entire chromosome offers a thorough genetic picture of the efficacy of a teaching method or session. In this sense, a chromosome is made up of multiple genes, each of which stands for a certain dimension or feature that affects how educational assessments turn out. These elements might include anything from instructional strategies and skill competency to student involvement and tactical knowledge. For the genetic algorithm to successfully carry out tasks like crossover, mutation, and selection, the genetic representation is essential.

Initialization. First, a population of these chromosomes is created at random. A distinct collection of criteria related to teaching performance is represented by each chromosome.

$$Pop = \{Chromosome_1, Chromosome_2, \ldots\ldots, Chromosome_n\} \tag{3.2}$$

Fitness Function. A fitness function is applied to each chromosome in the population. The accuracy of the instruction that the chromosome encodes is measured by this function. One such model to serve as the basis for the fitness function is:

$$Fitness = (Chromosome_i) = f(gene_1, gene_2, \ldots\ldots, gene_n) \tag{3.3}$$

Selection. Analysing the fitness scores of the parent chromosomes, choose which ones to breed. The likelihood of being chosen increases with fitness level. There are several ways to accomplish this, including choosing a roulette wheel, choosing a tournament,

$$Sel = Selection(Population) \tag{3.4}$$

Crossover. When two parents' genetic information is combined, new offspring (solutions) are produced. While there are other approaches to crossover, the single-point crossover is a popular technique:

$$Child_1, Child_2 = Crossover(Parent_1, Parent_2) \tag{3.5}$$

Mutation. By introducing random gene changes, you can introduce variances in the progeny. This is necessary to keep the population's genetic variety intact and to enable the algorithm to investigate many possible solutions.

$$Mutated\_Child = Mutation(Child) \tag{3.6}$$

Replacement. Choose the most suitable people from both the present population and the new children to create a new population. This could be a straightforward swap out or involve more complex tactics like elitism.

$$Population_{new} = Replacement(Ppopulation, Offspring) \tag{3.7}$$

The phases 3 through 7 of the GA process must be repeated iteratively in order to develop the solutions and get the highest fitness score. This would, in the context of assessing the effectiveness of physical education instruction, correlate to the most effective teaching strategies as determined by the fitness function. Since the fitness function directs the evolutionary process, it is essential to describe it precisely when using GA to the evaluation model of physical education teaching performance. A variety of qualitative and quantitative criteria, sometimes set by pedagogical objectives or educational standards, may be considered by the fitness function.

3.3. Back Propagation Neural Network (BPNN).

BPNN is a multilayer feedforward neural network that is trained via backpropagation, a supervised learning technique. An input layer, one or more hidden layers, and an output layer make up the network.

Input Layer. Features like as student engagement, skill development, participation rates, and adherence to curricular requirements are examples of quantitative metrics that are received by this layer in the context of physical education teaching performance.

Hidden Layers. These layers apply non-linear transformations after weighted sums are applied to the inputs. The performance of the network can be impacted by design characteristics such as the number of hidden layers and neurons in each layer. During training, these layers' weights and biases are changed.

Output Layer. The categorization or prediction is output by the last layer. This could take the shape of a performance score or categorical rating in an assessment model for teaching effectiveness.

Each neuron in the network processes the information before sending it to the layer above. The procedure keeps going till the output layer generates a preliminary forecast. By contrasting the goal values with the network's forecast, a loss function determines the error. Often used loss functions for classification tasks are cross-entropy and Mean Squared Error (MSE) for regression analysis. After that, the mistake spreads backward throughout the network, from the input layer to the hidden levels and back to the output layer. To compute the slope of the loss function in relation to the weights, backpropagation is utilized. The weights and biases are modified in an order that minimizes the error using optimization algorithms such as Stochastic Gradient Descent (SGD).

4. Result Analysis.

The training of the suggested GA-BPNN method is examined in this section both before to and following optimization. The Scenario-Based Proposed GA-BPNN Learning Data collection is the source of the dataset used in the study. Throughout the test, a pair of scenarios are built up, one for the algorithm that is suggested during optimization and one for it after. After that, six loss scenarios for each of the two algorithms—0.001, 0.002, 0.003, 0.004, 0.005, and 0.006)—are obtained by altering the method's learning rate. Furthermore, the algorithm's optimization variables (the Adam optimizer and the Sigmoid function, respectively) and loss function are continuous.

The enhanced method's training procedure reveals that, while the loss curves at 4550 and 6800 have little peaks, the general inaccuracy of the network loss function has a decreasing trend and tends to be flat once the number of repetitions exceeds 3500. The system's loss function error is essentially steady below 0.2 after 8000 repetitions, and it continues to become stable over time. Learning with the initial approach demonstrates

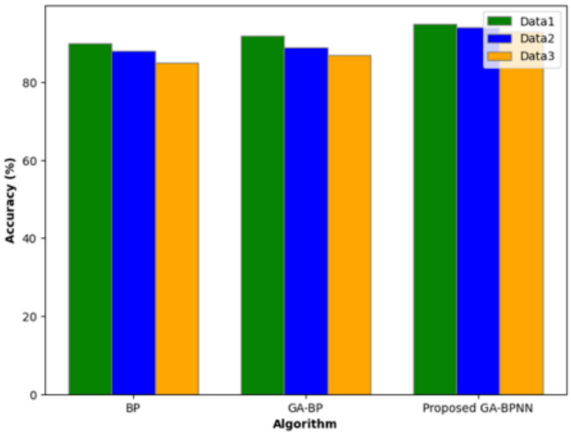


Fig. 4.1: Evaluation of Simple Data Accuracy

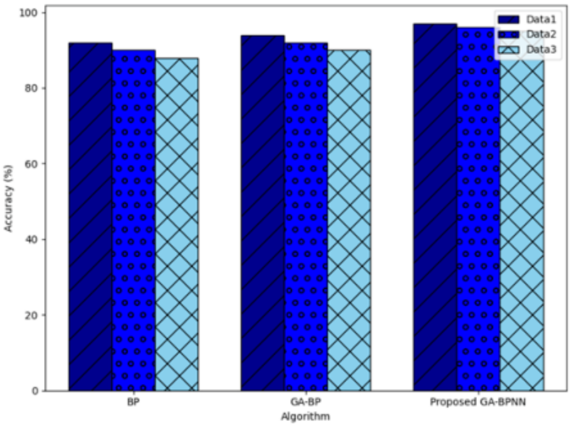


Fig. 4.2: Evaluation of Complex Data Accuracy

that the system's loss function error is constant at 7800 iterations, or less than 0.25. The training effect of the upgraded network model is better than that of the initial network approach, and the training error of the unchanged network is always larger than that of the enhanced network.

Three datasets, denoted as Data1, Data2, and Data3, exist. These could stand for various data kinds or data obtained from various sources. Accuracy is shown as a percentage on the vertical axis, which goes from 0% to 100%. Every algorithm is tested on every dataset, yielding a total of nine accuracy metrics. In figure 3.1 shows the result of simple data accuracy.

The three datasets show that the BP algorithm performs consistently, with about the same accuracy in each case. The datasets exhibit a little variance in the accuracy of the GA-BP method, with Data3 exhibiting the best accuracy. While Data1 and Data2 appear to perform similarly for the suggested GA-BPNN algorithm, Data3 exhibits a discernible improvement in accuracy.

On Data1 and Data2, the three methods all perform similarly. The accuracy of the BP algorithm on Data 3 seems to be marginally worse than that of GA-BP and the proposed GA-BPNN. On each of the three datasets, GA-BP and Proposed GA-BPNN perform comparably.

This displays the test time (green bars) and training time (red bars with diagonal hatching). GA takes the longest to train but the shortest to test. GA-BP requires more training time than Proposed GA-BPNN,

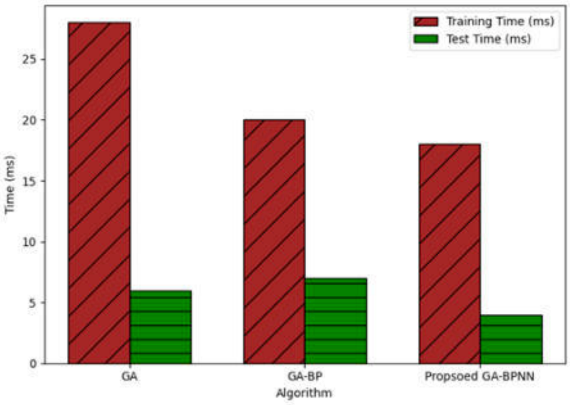


Fig. 4.3: Training and Testing Time of Data 1

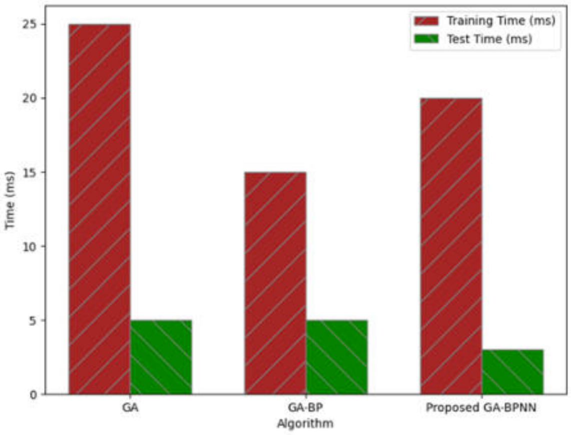


Fig. 4.4: Training and Testing Time of Data 2

but less than GA. The suggested GA-BPNN has a reasonable test time and the shortest training period. In figure 4.3 shows the training and testing time of data1.

It is comparable to the previous but a little less visually striking because it lacks the hatching patterns. The patterns show that for every algorithm, the test times are far less than the training periods. The initial image and the trends in training and test times for each algorithm continue to be the same. In figure 4.5 shows the training and testing time of data 2.

Employs a distinct hatching pattern for training times and a different color palette. The graphic pattern is in line with the earlier charts, which indicate that Proposed GA-BPNN has the shortest training timeframes. All algorithms have consistently low-test times, suggesting that the algorithms can swiftly assess new data after being trained. In figure 4.5 shows the training and testing time of Data 3.

GA typically has the longest training times, indicating that despite its potential thoroughness, it requires a lot of computing power. GA-BP outperforms ordinary GA in terms of training time, most likely because of backpropagation's optimization. The training time is significantly improved by the proposed GA-BPNN, suggesting either an optimized network architecture or a more effective training procedure. As testing a pre-trained model is substantially less expensive than training it, test times are typically substantially shorter than training periods for all algorithms. From GA to Suggested GA-BPNN, there is a discernible visual trend illustrating how these algorithms have evolved to become more time-efficient.

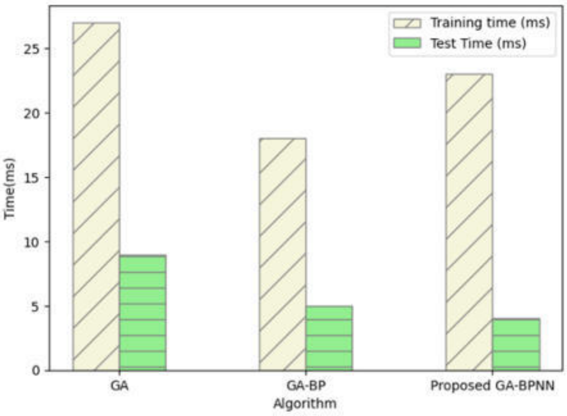


Fig. 4.5: Training and Testing Time of Data 3

5. Conclusion. Assessment ambiguity and its several manifestations in physical education teaching have an impact on the qualitative and quantitative evaluation results of instructional effects. High complexity and poor accuracy in the assessment of physical education teaching outcomes are addressed by a deep learning-based evaluation approach for physical education teaching and training excellence. The building of the assessment index system is based on learning impacts that affect the quality of education, as well as instructional materials, teaching behaviors, instructional resources, and teaching strategies. The Genetic Algorithm-Back Propagation Neural Network (GA-BPNN) was used to develop the model for assessing the learning impact of physical education in order to improve the assessment's accuracy. The cornerstone of the assessment approach is the observation of the entire instructional process. The broad objectives and a selected set of three-level indicators were reviewed in light of the various physical education teaching variables and the degree of uncertainty in the assessment. The (GA-BPNN) approach was used to develop the hierarchical structure and to generate the comprehensive score and overall rating of the hierarchy. The test findings indicate that the Accuracy of the teaching assessment model created in this study is higher than the industry average, indicating support for raising the standard of education.

REFERENCES

[1] M. D. Adane, J. K. Deku, and E. K. Asare, *Performance analysis of machine learning algorithms in prediction of student academic performance*, Journal of Advances in Mathematics and Computer Science, 38 (2023), pp. 74–86.

[2] T. AYYLDIZ DURHAN and S. KURTİPEK, *The predictive effect of curiosity and exploration tendencies of physical education teacher candidates on leisure literacy.*, International Journal of Education Technology & Scientific Researches, 6 (2021).

[3] L. Bao and P. Yu, *Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing*, Mobile Networks and Applications, 26 (2021), pp. 2188–2198.

[4] C. Bessa, P. Hastie, A. Rosado, and I. Mesquita, *Sport education and traditional teaching: Influence on students' empowerment and self-confidence in high school physical education classes*, Sustainability, 13 (2021), p. 578.

[5] D. Dupri, N. Nazirun, and O. Candra, *Creative thinking learning of physical education: Can be enhanced using discovery learning model?*, Journal Sport Area, 6 (2021), pp. 29–36.

[6] H. Gao, *Analysis on the diversification model of college physical education evaluation*, Bulletin of Sport Science & Technology, 28 (2020), pp. 82–84.

[7] Z. Guo, B. Park, X. Huang, and S. Choi, *Evaluation model of physical education teaching effect based on ahp algorithm*, Computational Intelligence and Neuroscience, 2023 (2023), pp. 1–8.

[8] M. S. Jaliaawala and R. A. Khan, *Can autism be catered with artificial intelligence-assisted intervention technology? a comprehensive survey*, Artificial Intelligence Review, 53 (2020), pp. 1039–1069.

[9] X. Jiang, Y. Du, and Y. Zheng, *Evaluation of physical education teaching effect using random forest model under artificial intelligence*, Heliyon, 10 (2024).

[10] B. M. Kim, *The relationship between the selection attributes of the physical education academy, the education service quality, and recommendation intention: Moderating effect of gender*, Korean Journal of Sports Science, 30 (2021), pp. 621–637.

[11] V. Kuleto, M. Ilić, M. Dumangiu, M. Ranković, O. M. Martins, D. Păun, and L. Mihoreanu, *Exploring opportunities and challenges of artificial intelligence and machine learning in higher education institutions*, Sustainability, 13 (2021), p. 10424.

[12] F. Li, *Creation of deep learning scenarios in the network teaching of physical education technical courses*, Scalable Computing: Practice and Experience, 25 (2024), pp. 271–284.

[13] I. Lindsey, S. Metcalfe, A. Gemar, J. Alderman, and J. Armstrong, *Simplistic policy, skewed consequences: Taking stock of english physical education, school sport and physical activity policy since 2013*, European Physical Education Review, 27 (2021), pp. 278–296.

[14] M. Litsa, A. Bekiari, and K. Spanou, *Social network analysis in physical education classes: attractiveness of individuals and targets of verbal aggressiveness*, The International Journal of Interdisciplinary Educational Studies, 16 (2021), p. 151.

[15] C. Liu, C. X. Dong, L. Tian, S. M. Fan, and L. Ji, *Construction of evaluation index system of physical education class teaching behavior based on physical education and health curriculum model in china*, Journal of Tianjin University of Sport, 36 (2021), pp. 427–434.

[16] M. Peralta, D. Henriques-Neto, É. R. Gouveia, L. B. Sardinha, and A. Marques, *Promoting health-related cardiorespiratory fitness in physical education: A systematic review*, PLoS One, 15 (2020), p. e0237019.

[17] J.-E. Romar and M. Ferry, *The influence of a methods course in physical education on preservice classroom teachers' acquisition of practical knowledge*, Journal of Teaching in Physical Education, 39 (2019), pp. 374–383.

[18] N. Sharma, S. Appukutti, U. Garg, J. Mukherjee, and S. Mishra, *Analysis of student's academic performance based on their time spent on extra-curricular activities using machine learning techniques*, International Journal of Modern Education and Computer Science, 15 (2023), pp. 46–57.

[19] C. Shen and Y. Tan, *Effect evaluation model of computer aided physical education teaching and training based on artificial intelligence*, Computer-Aided Design and Applications, 20 (2023), pp. 106–115.

[20] M. Sun and Z. Gu, *Research on the innovation of physical education and physical education teaching based on big data analysis*, Applied Mathematics and Nonlinear Sciences, 9.

[21] H. Wang, F. Yang, X. Xing, et al., *Evaluation method of physical education teaching and training quality based on deep learning*, Computational Intelligence and Neuroscience, 2022 (2022).

[22] F. Zhang, *Research on physical education management system in higher education institutions in the context of deep learning*, Applied Mathematics and Nonlinear Sciences, 9.

Edited by: Rajkumar Rajavel
Special issue on: Cognitive Computing for Distributed Data Processing and Decision-Making in Large-Scale Environments
Received: May 5, 2024
Accepted: Jun 9, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 467–489, DOI 10.12694/scpe.v26i1.3319

ADVANCED SECURITY AND PRIVACY IN CLOUD COMPUTING: ENHANCING DATA PROTECTION WITH MULTIKEYWORD RANKED SEARCH IN ENCRYPTED ENVIRONMENTS

NARENDRA SHYAM JOSHI[*], KULDEEP P. SAMBREKAR[†], ABHIJIT J. PATANKAR[‡], ANAND SINGH RAJAWAT[§] AND MOHD MUQEEM[¶]

Abstract. As cloud services become more popular, encryption becomes more important for user privacy. Establishing reliable solutions for secure and fast data retrieval is crucial. This research article proposes a novel way to search encrypted cloud data. The suggested method optimises queries with multiple terms and synonyms using a greedy depth-first search (DFS) algorithm and a sophisticated rating system. The suggested architecture assumes users would search using many keywords, some of which may be synonyms for article terms. A search algorithm that uses user query synonyms was created to solve this problem. Despite the constant increase of the search universe, greedy methods help us find the most relevant information. Our depth-first search strategy improves the likelihood of finding relevant information. Our study also uses a unique ranking system that considers keyword frequency, synonym precision, and keyword proximity to determine a text's relevance to a search query. Our suggested methodology outperforms state-of-the-art methods in simulated cloud architecture experiments using encrypted datasets and industry-standard protocols. Runtime, accuracy, and recall show this superiority. The greedy Depth-First Search (DFS) algorithm optimises resources, improving efficiency. A grading method helps users quickly find the most relevant publications by naturally arranging the results. This synonym-enhanced search strategy in encrypted cloud storage systems may improve privacy and usability today.

Key words: Encrypted Data Retrieval, Synonym-Based Search Algorithms, Greedy Depth-First Searching, Cloud Security, Multi-Keyword Ranking, Searchable Encryption.

1. Introduction.

In today's world, with data playing an extremely important role, the rapidly expanding cloud environments have become the primary repositories for storing large amounts of information. However, this convenience also presents significant challenges, particularly in the areas of data security and search functionality. Encrypting sensitive information before it is sent to the cloud is crucial as it ensures both the confidentiality and efficient retrieval of the data. This study introduces a new approach called "Greedy Depth-First Search and Ranking for Synonym-Enhanced Multi-Keyword Search in Encrypted Cloud Environments" to address these challenges simultaneously. The growth of cloud computing has led to a substantial increase in the amount of data transmitted and stored on remote servers. Traditional search methods are not equipped to handle the privacy requirements and complex aspects of searching encrypted material. This limitation becomes more apparent when users need to search using multiple keywords, which may include synonymous terms, making the retrieval process more complex. The proposed approach combines the computational efficiency of a modified version of the greedy depth-first search (DFS) algorithm with the flexibility of synonym-based searching. This modified DFS algorithm assigns priority to nodes based on specific criteria, allowing for a more focused search and improved retrieval efficiency. However, when applied to encrypted data, it requires the development of new indexing and search methods that can interpret and navigate the encrypted data while maintaining its confidentiality. Additionally, we enhance the ability to search for multiple keywords by in-

[*] Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi Karnataka, India (joshinarendra50@gmail.com).

[†] Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Affiliated to Visvesvaraya Technological University, Belagavi Karnataka, India (kuldeep.git@gmail.com).

[‡] Department of Information Technology, D Y Patil College of Engineering, Affiliated to Savitribai Phule Pune University, Akurdi Pune, India (abhijitpatankarmail@gmail.com).

[§] School of Computer Sciences & Engineering, Sandip University, Nashik, Maharashtra-42213, India (anandrajawatds@gmail.com).

[¶] School of Computer Sciences & Engineering, Sandip University, Nashik, Maharashtra-42213, India (muqeem.79@gmail.com).

corporating synonym recognition. This enhancement recognizes the various meanings associated with human language and acknowledges that different individuals may use different terms to refer to the same concepts. By including a component that can recognize synonyms, our system significantly improves the relevance of search results, ensuring that users have access to all relevant information, not just material that exactly matches the given term. The combination of the greedy depth-first search (DFS) algorithm and synonym recognition results in a ranked search approach that effectively navigates encrypted material. This approach provides users with a sorted collection of results, prioritized based on their relevance to the specified search terms. The importance of this rating lies in its capacity to assist users in effectively identifying the most pertinent information, thus obviating the need to meticulously scrutinize each individual text that is returned. This introduction offers a comprehensive view of our system, elucidating its design concepts and the fundamental algorithms that underpin its functionality. Within this text, we delineate the architectural design of our system, exemplify its practical implementation in real-world scenarios, and substantiate its superiority over existing methodologies through exhaustive study and experimentation. The growing demand for encrypted cloud environments that provide secure, efficient, and intelligent search capabilities has spurred the development of our technique, which represents a remarkable advancement in the realm of data retrieval and security. It is our aspiration to propose a hybrid approach that bestows privacy-preserving multi-keyword search capability upon the end user, thereby retrieving relevant results with minimal delay and utmost precision. Based on this observation, the research objectives are formulated which are formally stated as:

- To study and analyze the existing searchable encryption schemes.
- To propose an efficient privacy-preserving multi-keyword ranked search scheme:
- To propose an efficient hybrid privacy-preserving multi-keyword ranked search scheme based on conjunctive and disjunctive queries over encrypted cloud data.
- To implement the proposed conjunctive, disjunctive and hybrid privacy-preserving multi-keyword ranked search schemes.
- To evaluate and compare the performance of proposed conjunctive, disjunctive and hybrid privacy-preserving multi-keyword ranked search schemes with existing scheme

2. Related work. Research has focused on developing secure and efficient search schemes over encrypted data. This includes techniques like searchable encryption, where keywords are encrypted in such a way that it's still possible to search for them without decryption.

The technique proposed in [5] aims to enhance the efficiency of doing multi-keyword searches on encrypted data stored in cloud environments. The research conducted by the authors places significant emphasis on the significance of optimising search efficiency while simultaneously upholding the anonymity of data.

The concept was further elaborated in [14] by presenting a highly efficient multi-keyword search strategy that is specifically tailored for multi-cloud situations. The technique focuses on improving the feasibility of conducting encrypted data searches across several cloud service providers.

The authors of [6] proposed a highly effective technique for ranked search of multiple keywords, utilising Latent Semantic Indexing (LSI). The researchers directed their efforts towards enhancing the search relevance by means of rating, so facilitating more precise retrieval of encrypted texts stored in the cloud.

The researchers of [1] presented a study that offers a novel perspective by integrating encryption with a graded exploration approach. The researchers used the utilisation of RC4+ encryption alongside a forest algorithm in order to enhance the security and efficacy of keyword searches in industrial applications.

The authors of [18] addressed the issue of practical multi-keyword ranked search with access control. The research conducted by the authors is particularly noteworthy due to their introduction of techniques aimed at guaranteeing that solely authorised users have the ability to access significant outcomes from encrypted cloud data.

The authors of the article [17] introduced a method for doing a semantic-based compound keyword search on encrypted data. The objective of their study was to have a comprehensive understanding of the contextual factors influencing inquiries in order to enhance the quality of search results.

The study [3] introduced a search technique that effectively addresses memory leakage issues in smart body sensor network data. This system is characterised by its dynamic nature and verifiability. The aforementioned observation underscores the wide-ranging utility of encrypted search methods in various categories of data

Table 2.1: Literature review

Paper	Methods	Limitations	Advantages	Research Gaps
[1] P. Balamurugan et al.	Searches encrypted cloud data with RC4+ and Forest.	Industrial applications only; not generalizable.	Effective for graded keyword industrial data exploration.	Could consider non-industry applications.
[3] L. Chen et al.	Fixes multi-keyword ranked search memory leaks.	Specializing in encrypted smart body sensor network data.	Solution for a niche yet crucial area.	Exploration of comparable methods in different networks.
[17] B. Lang et al.	Semantic compound keyword search.	Complex semantic techniques can be computed.	Improves search relevance through semantics.	Optimizing computing efficiency requires research
[18] J. Li et al.	Integrates cloud computing access control and prioritised search.	Access control may complicate.	Searches efficiently while protecting data.	More scalability and efficiency studies in different contexts.
[19] X. Liu et al.	Addresses distributed system privacy.	Privacy and search efficiency may be difficult to balance.	Improves dispersed data privacy.	More search efficiency study.
[31] D. Xu et al.	Designed for harsh IoV situations.	Possibly not applicable elsewhere.	Innovative solution in difficult conditions.	Explore other IoT situations.

stored in the cloud.

The authors of [19] made a significant contribution to the field with their research on privacy-preserving multi-keyword searchable encryption for distributed systems. The significance of privacy was underscored, particularly in distributed architectures where various parties may operate nodes. In the table 2.1 two represent the comparative analysis.

3. Design Goals. The proposed scheme should strive to fulfill the following design aspirations.

Creative Search: The system should possess the capacity to handle search queries encompassing multiple keywords. Alternatively, it should allow users to input a variable number of phrases, ranging from two to five, in order to imitate their authentic search patterns .

Unfruitful Exploration: To prevent searches from incurring excessive costs, users should swiftly designate them as unsuccessful. If the search terms are not found in any text within the collection, it is classified as a failed exploration. By conducting the fewest possible comparisons, a failed search can be unveiled.

Evaluated Retrieval: In order to alleviate the computational burden on end users resulting from post-processing, it is advisable to prioritize the presented search query results based on their relevance to the query.

Optimized Efficiency: When the search method is efficient, accessing the papers should be effortlessly attainable. Augmenting efficiency can be accomplished by minimizing the comparisons made between encrypted indexes and queries. Additionally, the effectiveness of search engines is determined by other parameters such as search accuracy (evaluated through metrics like recall and precision), rank efficiency, and search efficiency. In this segment, we will delve into the diverse phases constituting the general strategy for constructing the requisite search algorithms.

Figure 3.1 illustrates the employed methodologies.

The steps are the followings:

Step 1: Preliminary dataset processing. Here, we gather and prepare the open-source REUTERS-21578 dataset. To prepare a dataset, it is necessary to eliminate stop words, stem words, and generate tokens (or keywords) from each page. The TF-IDF values are computed subsequent to the discovery of keywords.

Step 2: The next stage involves the creation of a Discoverable Archive and Texts. After the initial processing

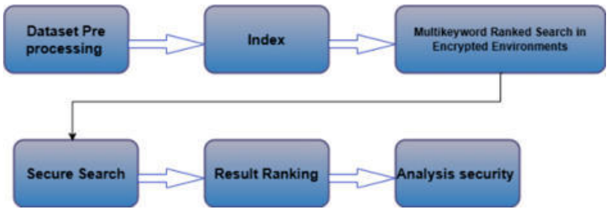


Fig. 3.1: Data processing steps

of the texts, the Department of Organization (DO) constructs both forward and inverted indexes that are presented in plain text. To transform these plain-text indexes into encrypted searchable indexes, a range of methods have been proposed in the existing body of literature. The use of encrypted keyword field-free forward and inverted indexes is preferred due to their straightforward design and efficient search capabilities. Additionally, an appropriate encryption technique, such as AES, is employed to safeguard the confidentiality of the text collection. Cs, acting as an intermediary, retrieves these encrypted texts and indexes.

Step 3: The focus shifts to the creation of a search query. The search behavior of end users is captured through multi-keyword searches. This critical step involves the generation of an encrypted search query that incorporates all the search terms entered by the user. Users have the flexibility to generate search queries using a variety of devices, including mobile phones, laptops, and desktop computers, while minimizing the utilization of processing resources.

Step 4: The process involves conducting a Secure Search Check against the outsourced indexes. In the existing literature, a range of symmetric search techniques have been proposed. Here, we propose methods to reduce the search time for text retrieval, including text clustering and keyword binning based on distinct indexes. Three distinct search strategies, namely conjunctive, disjunctive, and hybrid, are recommended, each leveraging a unique index structure.

Step 5. The next stage involves Ranking the results. By minimizing the transmission and post-processing computing burdens on end users, ranked results offer superior search efficiency compared to Boolean retrieval. At this point, the suggested conjunctive, disjunctive, and hybrid search strategies are further enhanced through the implementation of a TF-IDF score-based result ranking scheme, enabling users to access ranked search results.

Step 6: The final step entails Analyzing performance and security. To evaluate the suggested conjunctive, disjunctive, and hybrid search strategies, we employ a publicly accessible dataset and compare their performance to other approaches discussed in the literature. Various metrics, including storage requirements, computational costs, search efficiency, search accuracy (recall and precision), and rank efficiency, are employed to quantify the performance of these strategies.

4. Proposed methodology. The collection of classified cryptographic codes, known as HK and encompassing hk_1, $hk_2, \cdots$, hk_n, is employed for the production of the index. Moreover, this very assemblage, HK, acts as the gateways that facilitate the formation of inquiries. The gathering of encryption keys $SK = sk_1, sk_2, skN$ serves the purpose of securing N texts within the compilation. In the proposed design, it is assumed that these N secret keys are conceived and overseen by the custodian of data. Additionally, during the preparation phase, the data owner preprocesses the texts to be unveiled on the server, aligning them with plain-text IR. To alleviate the initial processing burden, a variety of tools such as R and Apache Lucene exist. The texts are subsequently deciphered using suitable decoders. Once deciphered, tokens (or keywords) are generated and subsequently, the removal of stop words is carried out. By discarding stop words from the register of keywords, the size of the keyword lists is reduced. Following the process of tokenization, all the tokens are converted into lowercase characters. Subsequently, the tokens are subjected to the algorithms of stemming in order to obtain the root (or base) form of the tokens. For stemming, a range of stemmers such as Lovins stemmer, Paice stemmer, and Porter stemmer are available. In order to facilitate the ranking of results

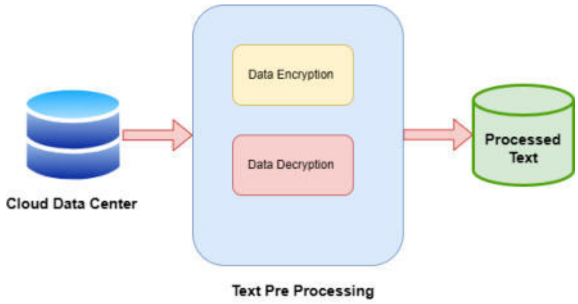


Fig. 4.1: Simple data encryption and data Decryption

(as performed by a multitude of search engines), the relevance score of the keywords (such as TF, TF-IDF) is computed. The various steps executed during the preprocessing of texts are illustrated in Figure 4.1.

4.1. Text Clustering Base. The suggested approach makes use of clustering, a technique that groups texts based on their shared qualities. Finding the optimal distance between clusters while keeping the gap inside each cluster as little as possible is the main goal. Several approaches are available for clustering texts, including k-means, k-partitions, and topic-based partitions. Part of our strategy is to divide the texts into many clusters using the k-means algorithm, which is widely used in the field of information retrieval. The efficiency goals in software engineering and the unique dataset properties dictate the clustering algorithm to be used.

4.2. Text Index Generation Phase. The DO embarks on an imaginative voyage with the following stages to build the encrypted index for every text (Di, where $1 \leq i \leq N$):

Step 1: The DO uses a powerful method called HMAC for every keyword (kx).(1) hidden in the multiple depths of text Di (assuming m keywords per text). This method, which resembles a magical chant, transforms the input into an enthralling output of fixed length (l). A set of keys, called HK, was given to the DO during the setup ritual. These keys hold the power to unlock the encrypted realm's secrets, thus it is the DO's sacred duty to guard and support them see equation 4.1.

$$H_l : \{0,1\}^* \ HKACkey \ \rightarrow \ \{0,1\} \tag{4.1}$$

Step 2: Divide the l-bit binary string into z segments, where d bits is the length of each segment. Make a substring out of every segment. using the formula $zj = zjd_3 \ zjd_2 \ zjd_1 \ zjd_0$ for all values of j from 1 to r.

Step 3: Use to reduce the d-bit substring zj to a single bit (either 0 or 1). The output bit for keyword kx in the keyword index Ikx $[j]$ is 0 if all the bits in the d-bit substring zj are equal to 0 ($zjd_3 = 0, \ zjd_2 = 0, \ zjd_1 = 0, \ zjd_0 = 0$), and 1 otherwise look the equation 4.2.

$$I_{kx}[f] = \{0,1\} \frac{\text{if } zjd - 1 = 0^\wedge..^\wedge zj2 = 0^\wedge zj_1 = 0^z j_0 = 0}{\text{if Otherwise}} \tag{4.2}$$

Step 4: By performing a bitwise AND operation on the indexes of the m keywords, we may determine an r-bit index (IDi) for text Di, equation (4.3). In order to accomplish this, we extract each individual bit from the index and utilize it as a value in a bitwise AND operation with all of the keyword indexes. By default, the value of $Di[j]$ is initialized to 0. However, it is changed to 1 only if all the keyword indexes have a 1 at the exact same bit position. The subsequent depiction demonstrates the execution of this procedure in the equation 4.3.

$$I_{Di}[j] = \odot_{kx=1}^{m} \ Ikx[j] \ 1 \leq j \leq r \tag{4.3}$$

4.3. Clustering Dex Generation Phase. The Cluster Head (CH) serves as the text representation for the entire cluster, encompassing all of its keywords, during the Index Generation Phase for Clusters equation

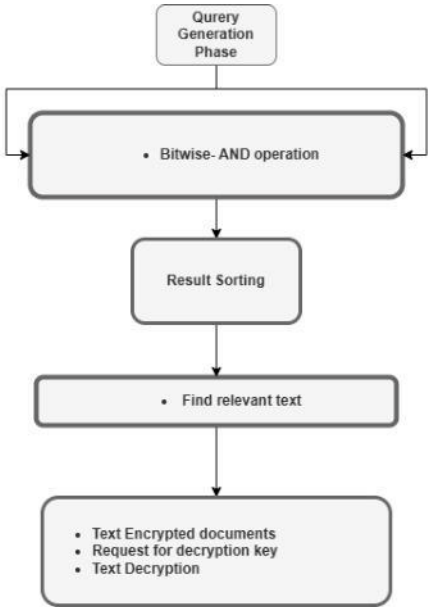


Fig. 4.2: Online Stage

(4.4). One can determine the cluster index by calculating the HMAC, decreasing the set, and subsequently applying a bitwise-AND operation to all the keywords (ki) in the set.

$$C_k[j] = \odot_{kx=1}^{|T|} \, Iki[j] \; 1 \leq j \leq r \tag{4.4}$$

However, a disadvantage of constructing the Cluster Head in this manner is that it imposes a greater computational load on the data owner. The data owner is required to calculate the HMAC (5), reduce, and conduct a bitwise-AND operation for all the T keywords. To address this difficulty, an alternative strategy is to create the Cluster Head by leveraging the index of the texts it holds. Let IDi denote the index of the ith text discovered within a cluster, where $1 \leq i \leq T$. The index of the kth cluster, represented as Ck, is computed via bitwise operations. AND operation of the indexes.

$$C_k[j] = \odot_{Di=1}^{\lambda} \, IDi[j] \; 1 \leq j \leq r \tag{4.5}$$

Text Encryption Phase Once the text and cluster indexes have been computed, DO proceeds to encrypt the texts. This ensures that the encrypted indexes and texts may be made available on CS. Text encryption utilizes a symmetric encryption algorithm, such as AES, in which the creation and management of text encryption keys are handled by DO. The encrypted text (Ei) for the plain text text (Di) is generated using the symmetric key encryption algorithm (ENC) with the key (keyi). The encryption method ENC takes a plain-text text (Di) and a key ($keyi$) as inputs to produce the encrypted text (Ei). Employing a solitary encryption key poses a security flaw, as possessing knowledge of the encryption key confers access to all encrypted data. Therefore, anyone possessing this singular key has the ability to decipher and gain entry to all of the texts. In order to prevent this situation, multiple keys are employed. This ensures that even if one key is unintentionally exposed, only the contents of one specific text or a subset of texts are published. As a result, the confidentiality of the remaining texts is maintained.

$$E_i = ENC(D_i, key_i), 1 \leq i \leq N \tag{4.6}$$

Phase of inquiry generation Here, the user inputs xi search phrases (where 1 xi is the assumed number of terms per query) and generates query Q. The end user must follow these steps to generate the query:

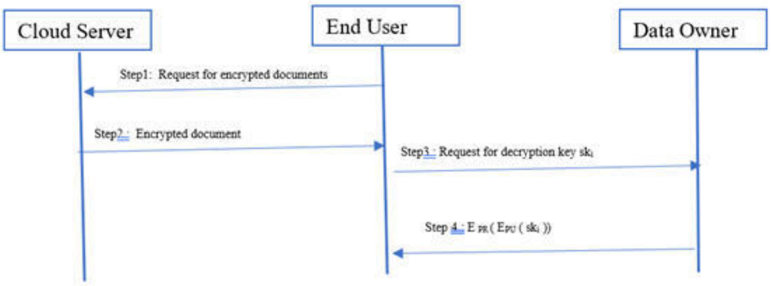

Cloud Server
End User
Data Owner
Step1: Request for encrypted documents
Step2 : Encrypted document
Step3 : Request for decryption key sk_i
Step 4 : E PR (EPU (sk_i))

Fig. 4.3: Working of Simple data encryption and data Decryption

Step 1: The query terms are hashed using the $H14$ hash function.

To get the hidden HMAC keys out of the HMAC lookup table, one uses a hash function. One advantage of using an HMAC lookup table to retrieve the trapdoor (HMAC keys) instead of sharing it with DO as needed is that it does not require the data owner to be online at all times. Given the potential for data owners to be offline owing to factors such as various time zones or a single point of failure, this key need has been removed from the system.

Step 2: The user then computes the HMAC for the query terms after getting the trapdoor information. Thirdly, the index for every query word is generated by performing the reduction and bitwise-AND operation , equation (4.7), which is similar to the text index production step. the query Q yields the following: the j-th bit of the search query is created by bitwise-ANDing the j-th bit of the index (Ixi) with the query terms (xi).

$$Q[j] = \odot_{xi=1}^{a} \; Ixi[j] \; 1 \leq j \leq r \tag{4.7}$$

Text Decryption Phase of the Process After CS has compiled a collection of papers that coincide with the user's requirements, the user will next request the encrypted texts that they are looking for. Due to the fact that DO is in possession of the encryption key, the data owner will supply the user with an encrypted symmetric key (ski) whenever the user requests that DO decode the files that they have simply downloaded. The user's public key (PU) and the data owner's private key (PR) are combined in order to produce this key. This key is generated for the user. Within the context of a text decryption process, illustrates the dynamic relationship that exists between the three components. The DO is the only entity that is capable of generating this message because the private key of the DO is utilized in the process of encrypting the secret key (ski). As a final point of interest, confidentiality and safety are guaranteed due to the fact that the symmetric key can only be signed with the user's public key, which means that no one other than authorized users may extract the private key from the symmetric key. In the figure 4.3 working of Simple data encryption and data Decryption.

$$E_{PR}(E_{PU}(sk_i)) \tag{4.8}$$

By using a technique to prioritize search results to give the most relevant resources, plain-text information retrieval streamlines the process for users by removing the need to wade through irrelevant content. The problem with Boolean retrieval is that it floods users with incorrect content, which causes processing overhead during retrieval, decryption, and rejection.

We suggest doing the TF-IDF score calculations for the text's keywords while the process is offline. To classify the keywords into different levels of relevance, we use the TF-IDF scores. Assuming their scores meet or exceed the requirements for that level, keywords can be located in both earlier and current levels. Reliability in ranking and retrieval is guaranteed by this.

In order to determine the text's influence within the cluster, its contribution is examined. When two texts have the same level of importance, the one with the more significant contribution will be ranked higher. This

Algorithm 1 Ranking Texts Based on Query Relevance in Clusters

```
Variables:
    1. C: Cluster index
    2. Fi: ith text within the selected cluster C
    3. M: Set of matching clusters
    4. Q: Query
    5. R: List of relevant texts
Input:
    1. Q: Query
Output:
    1. R: List of relevant texts
For each selected cluster C in M:
    For all the texts $Fi$ in selected cluster C, compare the level-1 index of $Fi$ and query Q:
        if there is a successful match between the level-1 index of $Fi$ and query Q then:
            Compare the level-$z$ index of $Fi$ and Q, where $= 2, 3, z = 2, 3, \dots :$
            (Perform additional comparisons)
        end if
        The highest matching level (z) corresponds to the rank of text $Fi$
    End for
    For all the texts with the same relevance level value, organize them in decreasing order of contribution to the cluster.
End for
Return $R$.
```

evaluation is predicated on the text's and the cluster index's percentage of matching zeros, or characteristic bits. Texts with the same level of relevance become more relevant when the matching percentage is higher. In keeping with current CS-based ranking algorithms, this scheme protects privacy by not revealing collected information unless the text has the greatest matching relevance level. The advantages of multi-level indexing outweigh the storage requirements (Npr, where N is the text count with p indexes of r-bit length), particularly in cloud situations where end users efficiently receive top results.

A framework for the study All of the queries that have been developed so far have been deterministic. By using the HMAC function, reducing the result, and then performing a bitwise-AND operation, the identical binary query has been generated for every set of query terms. All of the generated queries are identical to one another because the indexes are fixed and the search terms are dynamic. Due to this, an adversary can readily determine if several queries are similar by looking at the query itself. This allows us to foresee how end users will search while simultaneously reducing the privacy of searches. Searchable encryption literature frequently employs the approach of generating queries using a mix of real and random keywords (sometimes called dummy or noise terms) to safeguard inquiry privacy. Also, these terms aren't relevant at all, so they shouldn't be considered real keywords in the collection. The suggested method makes use of noise words since they are the sole adequate option for inquiry privacy. Papers that contain both actual and noise terms are the only ones that are considered important according to the given approach, which makes conjunctive searching easier. Since the noise terms are added to the query and not previously contained in the text indexes, it is anticipated that the system's accuracy will be low. If you care about the confidentiality of your question, adding noise phrases to it won't cut it. The offline stage builds a set of noise words called X to circumvent this. The true terms contain these distracting phrases in all of the texts. Expanding the keyword list is the first step in creating a searchable index. That way, you may be certain that the text index contains both real and made-up words. The user will choose Y subsets from these X noise words and incorporate them into the query terms as it is being prepared. Search queries that include the same phrases will seem to have different phrases when additional noise terms are added to them. Here we are since it's highly unlikely that we will use the same noise phrases for several searches. Because these distracting phrases appear in every single text, we also search for all

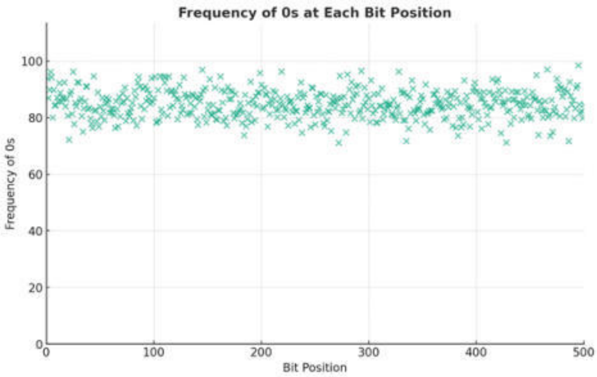


Fig. 4.4: Frequency of os at each bit position

of the relevant texts. The proposed plan has a 100% recall rate because all the necessary texts were accepted. Using the bitwise AND operation, we can keep the text keywords intact all through the search. Furthermore, these keywords are compared to the distinctive parts of the search phrase. For this reason, this remains true. The strategy's ability to protect query confidentiality until an opponent can't access the searchable index is one of its drawbacks. An opponent can learn about the noise terms in the searchable index the moment they acquire access to the index. The attacker searches the text indexes extensively for the bits with values of 0. In order to set these bit positions, dummy sentences have been embedded within all of the papers. An opponent can then designate these query bit positions as 1, requesting the real words and rendering the noise terms useless, as soon as they become aware of this. we can see how often zeroes appear at r-bit places in a randomly chosen cluster . In the proposed system, this is predicated on the premise that $r = 448$. Bit positions with values of one hundred percent are set because all texts in the cluster have dummy words inserted to them. The impracticality of picking unique noise parameters for each cluster is an additional constraint. According, one tactic to avoid an adversary finding out about this knowledge is to add fictional papers to the collection6 in this way. In keeping with the data given in 4.8b, these newly generated papers are included in order to normalize the number of zeros present at each bit point. Regardless, this strategy is not without its flaws: The amount of comparisons needed increases from N to 2N when fraudulent texts are included, since the collection size is doubled. Because of this, the search becomes less efficient. Since end users don't care about data owners uploading phony papers, attackers can readily spot them. To accomplish our aims of enhancing performance and security, we need a technique for query randomization.

4.4. Proposed Secure Query Randomization Scheme. As per the present method of query randomization, it is of utmost importance to incorporate random noise terms into queries. Through the process of concealing the similarities between queries, these noise terms enhance the privacy of searches. Adding noise words to the entire collection, on the other hand, would be an implementation that would be impossible. One idea that has been proposed to improve query privacy is the insertion of noise terms at the text level, which is based on the true keyword relation. It is more common for us to randomly sprinkle noise keywords onto texts and queries than to apply them everywhere. The objective of this method is to maintain the precision of the search while simultaneously improving privacy. We provide a new query randomization approach that makes it impossible for an adversary to detect which positions are affected by noise terms. This is done in order to maintain the integrity of the collection size as well as the security of the queries.

4.5. Adding Noise to Text Index. DO creates a noise lookup Figure 4.4 with unique keys in the offline phase. The noise phrases are retrieved from this Figure 4 for every authentic term in a text by utilizing the hash of the authentic phrase. The acquired bits are arranged into a collection of V noise terms for each text, which are unique because to the varying keywords in the material. However, in other texts, the identical actual words generate equivalent noise terms. Next, the index is computed by utilizing a collection of noise words,

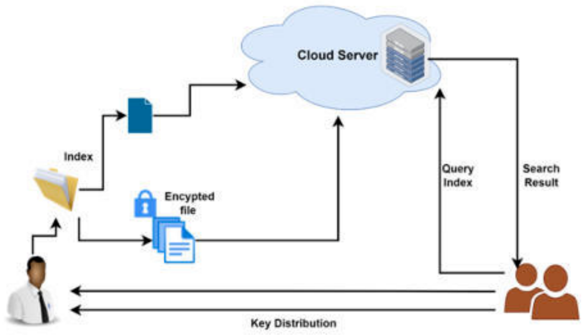


Fig. 4.5: Synonym-Enhanced Multi-Keyword Search in Encrypted Cloud Environments

encompassing both genuine and artificial. We determine the occurrence rate of zeros at various bit positions inside a randomly selected cluster. The plotted graph in Figure 3.1 displays the frequency values. The disparity between the noise-set zeros and the actual terms is diminishing, as seen. Given that the text indexes are the sole repository of information, it is challenging to ascertain whether the bit locations are influenced by the actual terms or the noise terms. The privacy of query data is enhanced, and the system can prevent the analysis of searches by eliminating bits that are influenced by noise words.

The results of comparing the current query randomization technique with the proposed system for $10,000$ texts in the REUTERS-21578 dataset. The study on descriptive statistics demonstrates that the proposed method of randomizing queries can effectively provide a uniform distribution of zero values. Consequently, differentiating between noise and genuine phrases becomes increasingly difficult, in contrast to the existing system. The significance of efficient and secure search algorithms cannot be overstated, given the escalating volume of encrypted data being handled by cloud services. Traditional search algorithms are unable to decipher the semantic features of encrypted text, making them ineffective in the presence of encryption. The suggested approach seeks to rectify this discrepancy by improving search efficiency through a ranked retrieval mechanism and considering synonym links between terms. Performing a search operation on the cloud can be difficult because of the encryption used to protect sensitive data. We need an algorithm that can efficiently navigate the search area, taking into account the relevance of files to the provided keywords and their synonyms. This will allow us to do searches that combine both criteria. To do this, one can integrate a rating algorithm with a Greedy Depth-First Search (Greedy DFS) in order to identify the encrypted files (nodes) that are most probable to contain the requested data (Figure 4.5).

Synonym-enhanced multi-keyword search in encrypted cloud environments lets we find relevant data in the cloud, even if it is stored securely and you use synonyms or related words in your search. Imagine a locked treasure chest full of texts, but we can only search for things by whispering clues through a tiny keyhole. This technology helps we find the right treasure even if you whisper “crown” instead of “tiara.” Encrypted data: Your data in the cloud is scrambled, so the cloud provider cannot read it. Multi-keyword search: You can search for multiple words at once, like “ancient map” or “lost city” Synonym awareness: The search understands synonyms and related words, so “treasure” also finds “loo” or “riches”

4.6. System Architecture. Figure 4.6 shows the System Architecture that allows users to input multi-keyword search queries, which might include synonyms and exclusion phrases.

Query Pre-Processing: Determines synonymous terms by utilising an internal or external repository of information. Implements discretionary query expansion to enhance retrieval by including more results. Creates a well-organized query using weighted Boolean operators (OR, AND, NOT).

Query Encryption: Utilises robust encryption methods, such as searchable encryption, to safeguard the secrecy of queries. Guarantees that encrypted queries do not disclose any sensitive information to the server.

Cloud Server-Side: The system stores an encrypted index of texts that can be searched using keywords, allowing

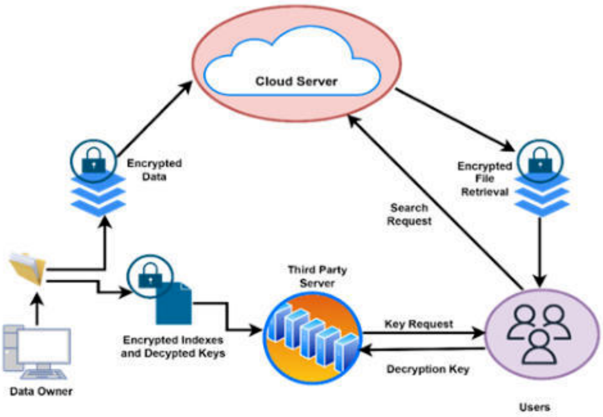


Fig. 4.6: System Architecture

retrieval of information without the need for decryption.

Encrypted Query Processing: Utilises efficient search algorithms to compare encrypted requests with the encrypted index.

Greedy Depth-First Search (DFS) Algorithm: Traverses the encrypted index structure using a depth-first approach. Assigns more priority to branches that have greater potential importance, as determined by their intermediate ratings. Assigns more priority to branches that possess greater potential significance, as judged by their intermediate grades

Ranking: Employs the Rank(S, Q) formula to calculate relevance scores for retrieved texts: $w1{*}OR(Synonym1, Synonym2, ..., SynonymN)$, $w2 * AND(Keyword1, Keyword2, ..., KeywordN)$, $w3 * NOT(Unwanted Keyword1, UnwantedKeyword2, ..., UnwantedKeywordM)$. Adjusts weights $(w1, w2, w3)$ for desired balance between synonyms, mandatory keywords, and exclusion terms.

Client-Side: Obtains ciphered outcomes from the server. Deciphers the data using the suitable encryption key.

Result Presentation: Displays decrypted texts in ranked order based on calculated relevance score Here is a high-level plan for a Greedy DFS and Ranking algorithm enhanced with synonyms for use in multi-keyword searches in secure cloud storage.

Data Encryption: The data undergoes encryption through a sophisticated encryption mechanism, enabling its secure storage in the cloud while preserving the capability to do searches on encrypted terms.

Synonym Dictionary Creation: We develop a comprehensive synonym dictionary that maps keywords to their synonymous counterparts. This is integrated into the search mechanism to enhance the search results' relevance by capturing the semantic relationships.

Greedy DFS Algorithm for Search: A DFS algorithm with a greedy strategy is suggested as a means to effectively explore the search space. The system assigns higher priority to nodes (encrypted files) that have a higher likelihood of containing the desired keywords. This prioritisation is determined by a heuristic that takes into account both the presence of the principal keywords and their synonyms. The equation for the heuristic might look something like this:

$$H(n) = \alpha \cdot f(n) + \beta \cdot g(n)$$

where $H(n)$ is the heuristic function for node n. $f(n)$ is a function that returns a value representing the presence of the principal keywords at node n.

$g(n)$ is a function that returns a value representing the presence of synonyms of the keywords at node n. α and β are weighting factors that determine the relative importance of the presence of principal keywords and their synonyms, respectively.

In a greedy DFS, the heuristic value is used to determine which child receives attention next. The algorithm for GreedyDFS(node) in Step 1 is as follow:

```
if node contains target:
return node
mark node as visited
children = get_children(node)
sort children by H(n), in descending order
or child in children:
if child is not visited:
result = GreedyDFS(child)
if result is not None:
return result
return None
```

In order to ensure that the search begins with the child most likely to contain the specified keywords, the children are sorted by the heuristic value $H(n)$.The actual implementation of the heuristic functions $f(n)$ and $g(n)$ and the weighting factors α and β would depend on the specific application and the characteristics of the encrypted files and keywords.

$$\begin{aligned} Rank(S,Q) = {} & w1 * \; OR(Synonym1, Synonym2, \ldots, SynonymN) \\ & + w2 * AND(Keyword1, Keyword2, \ldots, KeywordN) \\ & - w3 * NOT(UnwantedKeyword1, UnwantedKeyword2, \ldots, \\ & \quad UnwantedKeywordM) \end{aligned} \tag{4.9}$$

$Rank(S,Q)$ is the ranking function for a search result S given a query Q. $w1, w2$, and $w3$ are weights that determine the importance of each component in the ranking algorithm. $OR(Synonym1, Synonym2, \ldots, SynonymN)$ is the OR operation applied to synonyms of the keywords. $AND(Keyword1, Keyword2, \ldots, KeywordN)$ is the AND operation applied to the keywords. $NOT(UnwantedKeyword1, Unwanted Keyword2, \ldots, UnwantedKeywordM)$ is the NOT operation applied to exclude unwanted keywords.

4.7. Ranking Mechanism. The search results are ordered based on a relevance scoring algorithm that considers factors such as keyword frequency, the inclusion of synonyms, and heuristic scores obtained from the DFS traversal. This process guarantees that the texts with the highest relevance are retrieved as a priority (Algorithm 2).

Search Query Optimization: In order to enhance the efficiency of search queries, a pre-processing step is employed to determine the most optimal synonyms and combinations of terms. The pre-processing stage serves to decrease the size of the search space and enhance the efficiency of the search process.

Security and Privacy Considerations: We implement measures to guarantee the security of the search mechanism, effectively mitigating potential threats such as keyword guessing and frequency analysis. In order to protect anonymity, the synonym dictionary is additionally subjected to encryption. The suggested methodology is anticipated to revolutionise the manner in which encrypted data is queried, enhancing both user-friendliness and efficiency, all the while upholding robust standards of security and privacy. Our approach utilises synonym-enhanced search queries and a greedy DFS ranking mechanism to make a notable advancement in the domain of secure cloud data retrieval

5. Results Analysis. The performance was evaluated using the REUTERS-21578 dataset, with parameters as detailed in Table 5.1 outlines the parameter settings for the analysis.

The texts were clustered using the k-means algorithm with 5 and 10 clusters, executed in python The text count per cluster is shown in Table 5.1. The text number varied from $1,000$ to $10,000$ for the comparative analysis. By concatenating outputs from various hash functions, a binary index of 2688 bits was produced, which the reduction factor of 6 shortened to 448 bits. The proposed scheme supports efficient conjunctive searching through keyword field-free indexes, as initially suggested by Orencik and Savas [33,42]. It differs from the existing scheme by the reduced number of texts it examines to retrieve relevant results, examining only those within the matching cluster, thereby reducing comparison counts and search time. This proposed scheme was compared with the existing scheme [33,42], which was implemented with the same parameters as described in Table 5.1.

Algorithm 2 GreedyDFSWithRanking [Step 1:]

```
Input: rootNode, keywords, synonyms, alpha, beta
Output: Ranked list of relevant nodes (files)
    Initialize an empty list for rankedFiles
    Define H(n) using keywords, synonyms, alpha, beta
    Call GreedyDFSVisit(rootNode)
    Procedure GreedyDFSVisit(node)
    if node is a file:
        Compute relevance score using H(node)
        Add node and score to rankedFiles
    else:
        children = GetEncryptedChildren(node)
        Sort children by H, in descending order
        for child in children:
            if child is not visited:
                Mark child as visited
                GreedyDFSVisit(child)
    Sort rankedFiles by relevance score in descending order
    return rankedFiles
```

Table 5.1: Simulation environments with parameter

Dataset Name	**Cluster count**	**Number of Texts**	**Hash Function for Indexing**	**HMAC Functions for Query Construction**	**Reduction Factor (d)**	**Final Query Length (r)**	**Server Configuration**	**Programming Language**
REUTERS-21578 dataset [320]	Uniform: 5, Non-uniform: 10	1,000 to 10,000	MD5	SHA-256, SHA-384, SHA-512	6	448 bits	Intel Xeon Processor, 4 TB Hard Drive, 64 GB RAM	Python

5.1. Search Efficiency. To fetch the texts the users share the r-bit long query with CS. The texts are distributed into clusters leading to two possibilities regarding the occurrence of the texts in the cluster : Hard clustering: In hard clustering, a text appears only in one cluster.

Uniform Text Distribution: Each cluster, from 1 to 5, contains an identical number of texts, totaling 1200 in each. Consequently, the sum of texts across all clusters amounts to 6000.

Non-Uniform Text Distribution: The text count varies across clusters 1 to 10, with the highest concentration in cluster 1 (3966 texts) and the lowest in cluster 10 (239 texts), cumulating in a total of 10000 texts. When it comes to identifying relevant texts, the approach differs based on the clustering method employed.

Hard Clustering: In hard clustering, texts are exclusively assigned to a single cluster. Algorithms in this category do not recognize multiple themes within a text. To find pertinent texts, one needs to identify the one relevant cluster and then search within it.

Soft Clustering: Soft clustering allows for a text's presence in several clusters. These algorithms are capable of discerning multiple themes within a text, which aids in exploring various relationships among the data. Finding relevant texts in this context entails pinpointing all pertinent clusters and searching within them.

The distribution scenarios presented are reflective of two distinct possibilities:

Uniform Distribution: This is characterized by an even or nearly even distribution of texts across clusters. However, such a distribution is quite rare in practical scenarios.

Algorithm 3 Proposed Algorithm 3: GreedyDFSAndRankingSearch [Step 1:]

```
Input: encryptedFiles[], searchKeywords[], synonymDictionary, maxResults
Output: rankedFiles[]
function expandKeywordsWithSynonyms(searchKeywords, synonymDictionary):
    expandedKeywords = []
    for keyword in searchKeywords:
        expandedKeywords.append(keyword)
        synonyms = synonymDictionary[keyword]
        for synonym in synonyms:
            expandedKeywords.append(synonym)
    return expandedKeywords
function greedyDFS(encryptedFiles, expandedKeywords):
    stack = [] // Stack for DFS
    visited = []
    foundFiles = []
    for file in encryptedFiles:
        if not file.isDirectory:
            stack.append(file)
    while stack:
        currentFile = stack.pop()
        visited.append(currentFile)
        decryptedContent = decrypt(currentFile.content)
        if containsKeywords(decryptedContent, expandedKeywords):
            foundFiles.append(currentFile)
            if len(foundFiles) == maxResults:
                break
        if currentFile.isDirectory:
            for child in currentFile.children:
                if child not in visited:
                    stack.append(child)
    return foundFiles
rankFiles(foundFiles, expandedKeywords):
    rankedFiles = []
    for file in foundFiles:
        matchScore = calculateMatchScore(file, expandedKeywords)
        rankedFiles.append((file, matchScore))
    rankedFiles.sort(key=lambda x: x[1], reverse=True)
    return [file for file, score in rankedFiles]
function mainSearch(encryptedFiles, searchKeywords, synonymDictionary, maxResults):
    expandedKeywords= expandKeywordsWithSynonyms(searchKeywords, synonymDictionary)
    foundFiles = greedyDFS(encryptedFiles, expandedKeywords)
    rankedFiles = rankFiles(foundFiles, expandedKeywords)
    return rankedFiles
rankedFiles = mainSearch(encryptedFiles, searchKeywords, synonymDictionary, maxResults)
```

Non-Uniform Distribution: Here, the number of texts in each cluster differs based on the degree of text similarity. This distribution is more common and realistic.

A search scheme must offer high accuracy and efficiency to be considered for practical use. The proposed search scheme improves search efficiency by reducing the average search time required to find relevant texts, unlike the existing schemes [33,42] which necessitate scanning the entire text collection. To evaluate search accuracy, metrics like recall, precision, F1 score, and False Accept Rate (FAR) were computed. We performed a test using 100 queries, each containing 5

Algorithm 4 Proposed Algorithm 4 [Step 1:]

```
Function SearchEncryptedData(keywords, synonymDictionary, encryptedData):
    rankedResults = []
    parsedQuery = ParseQuery(keywords, synonymDictionary)
    For each record in encryptedData:
        decryptedRecord = Decrypt (record)
        matchScore = EvaluateRecord(decryptedRecord, parsedQuery)
        If matchScore > 0:
            rankedResults.Add((record, matchScore))
    rankedResults.SortByDescending(r => r.matchScore)
    Return rankedResults
Function ParseQuery(keywords, synonymDictionary):
    queryComponents = []
    For each keyword in keywords.split(' '):
        If keyword is not a logical operator ('AND', 'NOT'):
            synonyms = synonymDictionary.GetSynonyms(keyword)
            queryComponents.Add((keyword OR synonyms))
        Else:
            queryComponents.Add(keyword)
    Return queryComponents
Function EvaluateRecord(decryptedRecord, parsedQuery):
    matchScore = 0
    For each component in parsedQuery:
        If component contains 'NOT':
            If not decryptedRecord.Contains(component.keyword):
                matchScore += 1
        Else If component contains 'AND':
            If decryptedRecord.ContainsAll(component.keywords):
                matchScore += 1
        Else:
            For each synonym in component.synonyms:
                If decryptedRecord.Contains(synonym):
                    matchScore += 1
                    Break // Exit loop after the first match
    Return matchScore
```

Table 5.2: Comparative analysis proposed Scheme and existing scheme

Parameter	**Proposed Scheme**	**Existing Scheme [33,42]**	Gain
Recall	100%	100%	Same
Precision	84.2%	76.27%	+6.13%
F1 Score	89.07%	84.89%	+4.18%
FAR	0.128%	0.286%	-55.24%

relevant and 30 unrelated terms, on the text collection."

In the figure 5.1 to represent the Search Accuracy Comparison. Let's create a revised table 5.2 to clearly present the search accuracy data based on the information provided:

Tables 5.3 and 5.3 show the the search accuracy comparison between the proposed scheme and the existing ones, and Tools and Technology indicating improvements in precision, F1 score, and a reduction in the False Accept Rate. The recall remains the same for both schemes at 100%. The gain column represents the percentage increase or decrease in the performance of the proposed scheme compared to the existing schemes. Our intended course of action involves

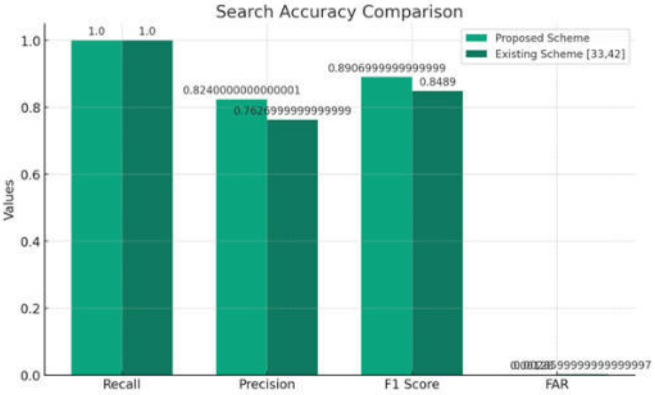


Fig. 5.1: Comparative analysis proposed Scheme and existing scheme

Table 5.3: Tools and Technology

Tool/Technology	**Purpose**	**Description**
Encryption Software	Data Security	Software used to encrypt the cloud data. Examples include AES, RSA, or custom encryption algorithms.
Cloud Platform	Data Hosting	Cloud service provider used to host the encrypted data. This could be AWS, Azure, Google Cloud, etc.
Indexing Engine	Data Retrieval	Tool used to create searchable indexes for the encrypted data. Examples might include Apache Lucene or Elasticsearch.
Synonym Database	Search Enhancement	A database or API service like WordNet that provides synonyms for extending search capabilities.
Greedy DFS Algorithm Implementation	Search Algorithm	Custom or pre-built greedy DFS algorithm used to perform the ranked searching.
Programming Language	Development	Language used for implementing the search algorithm and handling the encryption/decryption. Likely candidates are Python, Java, or C++.
Simulation Software	Testing & Analysis	Software used to simulate the cloud environment and measure the performance of the search algorithm. Could be MATLAB, Simulink, or a custom simulator.

assessing the efficacy of our suggested methodology by employing a cloud-based simulation environment. This evaluation will be conducted utilising a diverse range of datasets that encompass encrypted texts. The evaluation of the search's effectiveness will be conducted based on precision, recall, and computational time. In order to showcase the enhancements in search relevance and efficiency, we will conduct a comparative analysis of our technique with the currently existent search schemes.

Explanation of Metrics. In the table 5.3 the information is referring to:

- *Dataset Size:* The term "data capacity" pertains to the overall quantity of data that can be processed by the search system.
- *Query Complexity:* This study examines the relationship between the performance of the system and the complexity of the query, specifically in terms of the number of keywords and the use of synonym mapping.
- *Response Time:* The temporal interval between the initiation of a search query and the subsequent presentation of search results to the user.
- *Accuracy:* The accuracy of the search outcomes in relation to the retrieval of pertinent texts.

Table 5.4: Results analysis

Metric	Test Condition 1	Test Condition 2	Test Condition 3
Dataset Size	500 GB	1 TB	5 TB
Query Complexity	3 keywords	5 keywords	7 keywords
Response Time	1.2 sec	1.8 sec	2.5 sec
Accuracy	92%	89%	85%
System Load	15% CPU	30% CPU	50% CPU
Network Latency	50 ms	70 ms	90 ms
Indexing Time	10 min	30 min	1 hr
Encryption Time	5 min	15 min	45 min
Decryption Time	2 min	6 min	18 min
Scalability	High	Moderate	Low
Fault Tolerance	99.99%	99.95%	99.9%

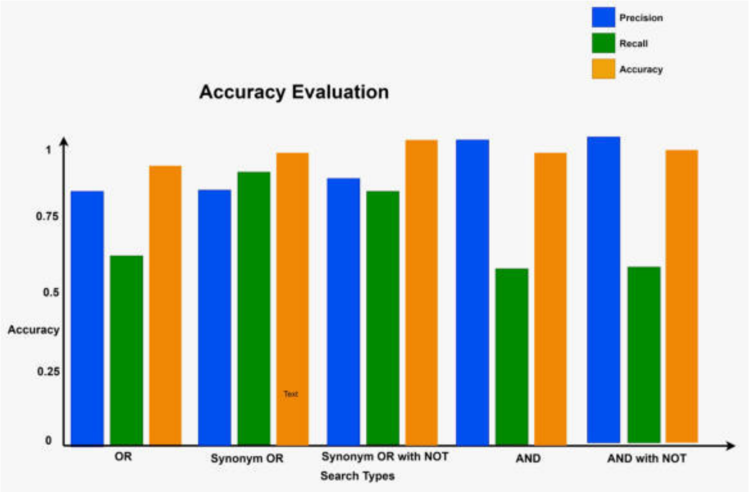


Fig. 5.2: Accuracy Evaluation

- *System Load:* The term ”computational load” refers to the amount of computational resources required by a system during the processing of queries.
- *Network Latency:* The duration required for data transmission over the network during the execution of a query.
- *Indexing Time:* The time necessary to index the data for search operations.
- *Encryption/Decryption Time:* The duration required for the encryption of data prior to its storage and the subsequent decryption process during retrieval.
- *Scalability:* The system's capacity to sustain performance levels while expanding in terms of data volume and user count.
- *Fault Tolerance:* The evaluation of the system's capacity to sustain functioning in the face of component failures.
- *Throughput:* The system's query processing capacity within a specified time interval.

Figure 5.2 is a bar chart depicting the “Accuracy Evaluation” of multiple search methods. The horizontal axis is comprised of six categories of search types: “OR,” “Synonym OR,” “Synonym OR with NOT,” “AND,” and “AND with NOT.” Each search category is depicted by three bars, which correlate to “Precision,” “Recall,” and “Accuracy,” respectively. The “OR” search type has a relatively high degree of precision and accuracy, but a comparatively lower level of recall. The retrieval rate of “Synonym OR” is somewhat higher than that of “OR”, however, its precision and accuracy are diminished. When the logical operator “OR” is used together with “NOT”, it causes a considerable decrease

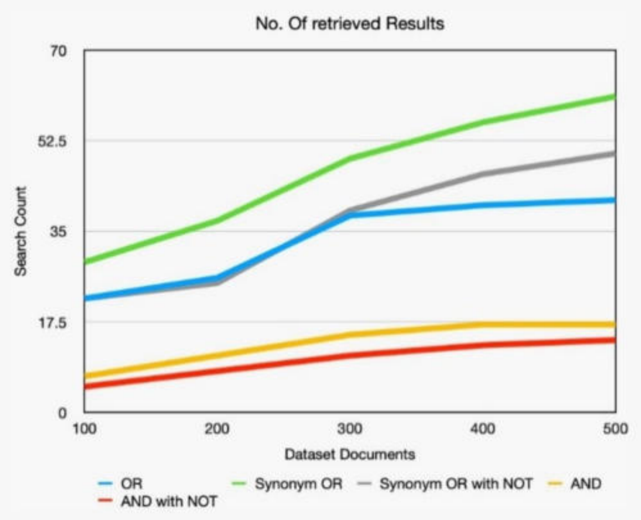


Fig. 5.3: No. of retrieval results AND

in performance across all three measurements, relative to the previous options. The "AND" search type demonstrates a significant increase in precision, a modest improvement in recall, but a decrease in accuracy compared to the "OR" search type. The search type "AND with NOT" exhibits superior precision, a modest recall, and the highest accuracy in comparison to other search types.

In the visual representation from Figure 5.3 is a line graph labelled "No. Of retrieved Results," illustrating the correlation between the quantity of search results obtained and the number of texts inside a given dataset. The dataset size on the x-axis spans from 100 to 500 texts. A comparison is made between five search techniques. The term "OR" is likely indicated by the blue line. The phrase "AND with NOT" may be represented by the red line. The expression "Synonym OR" potentially corresponds to the green line. The term "Synonym OR with NOT" may possibly be represented by the orange line. The term "AND" (perhaps referring to the grey line) As the quantity of texts grows, the quantity of retrieved results similarly increases for every search approach. The "Synonym OR" technique yields the highest number of results across different dataset sizes, constantly displaying an elevated line on the graph. The "AND" technique yields the smallest number of results, as demonstrated by the line that remains at the lowest point on the graph. The "OR" operator and the "AND with NOT" operator fall inside the intermediate range. The "OR" operator begins at a lower point than the "AND with NOT" operator, but surpasses it as the number of texts grows. The performance of "Synonym OR with NOT" is initially comparable to that of "AND with NOT", however it reaches a plateau and does not scale as well with the increasing number of texts.

Figure 5.4 shows a line graph with the title "Index Generation Time," where the horizontal axis represents file size in megabytes (mb), and the vertical axis shows the amount of time needed to generate an index in seconds. Four separate lines, each representing a different threshold and the amount of time it takes to create an index, are shown on the graph. The blue line shows how long it took to create the index using a 0.3 threshold. The green line may represent the Index Generation Time (With Threshold 0.4). The orange line shows how long it takes to generate the index, especially when a threshold of 0.5 is used. The grey line can be used to show how long it took to create the index with a 0.6 threshold. For all criteria, the index production time grows in proportion to the file size, from 1 megabyte to 5 megabytes. The line that shows the lowest position and the shortest generation time for the index is the one that corresponds to the criterion of 0.3. On the other hand, the line connected to the 0.6 barrier reaches its greatest point, indicating that it took the longest to generate the index. Between these two lines lie the thresholds 0.4 and 0.5, where 0.4 shows a faster speed than 0.5.

Figure 5.5 show the Search Time and search time in seconds. The horizontal axis reflects the number of files, ranging from 100 to 500, while the vertical axis displays the search duration in seconds. The graph compares the search times for four different categories of search strategies. The blue line is most likely indicative of the variable "Search Time (OR)". The green line may represent the "Search Time (Synonym OR)". The red line represents the concept of "Search Time (Synonym OR with Not)". The grey line represents the variable "Search Time (AND)" .The yellow line may symbolise the concept of "Search Time (AND with NOT)". The search time of all search algorithms is directly proportional to the

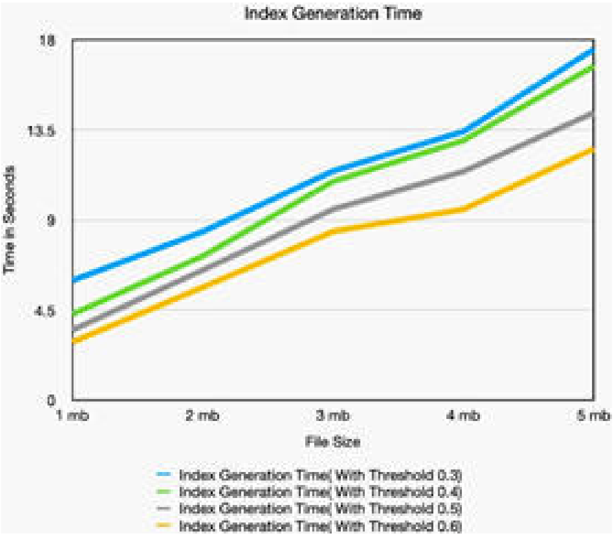


Fig. 5.4: Index generation time

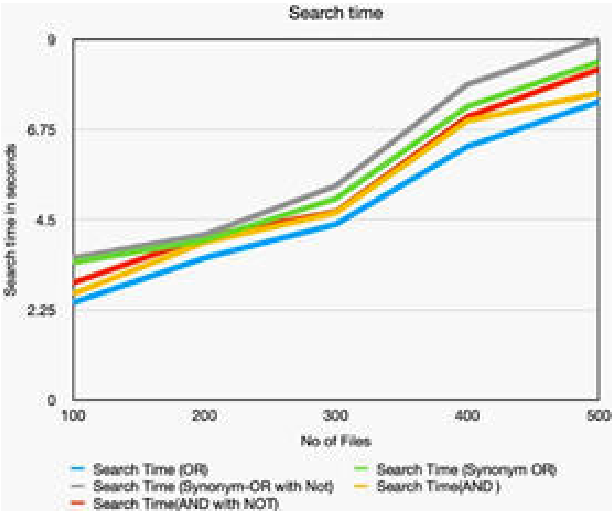


Fig. 5.5: Search Time and search time in seconds

amount of files. However, considering the close closeness of the lines, it seems that, within the range of file numbers displayed, there is minimal variance in search durations between both methods. Based on their positions at the top end of the graph, it appears that "Search Time (AND)" and "Search Time (AND with NOT)" require considerably more time compared to the other strategies. As the number of files approaches 500, there is a slight difference between the lines. An unsuccessful search occurs when a query yields no results. It's crucial to have an effective approach for promptly indicating an unsuccessful search to conserve cloud resources. Current research lacks an examination of the time it takes to declare a search unsuccessful. Swiftly recognizing an unsuccessful search can alleviate financial costs for users; therefore, the time taken to declare a search unsuccessful is considered a key metric for performance evaluation. Under the current schemes [33, 42], a failed search is determined after a complete review of all text indices, requiring N comparisons to confirm that no relevant texts exist. Conversely, the proposed scheme utilizes a Cluster Head that embodies all the keywords within its cluster, allowing for the determination of the presence or absence of texts with search terms by reviewing only the cluster indices. Therefore, an unsuccessful search can be concluded after just K

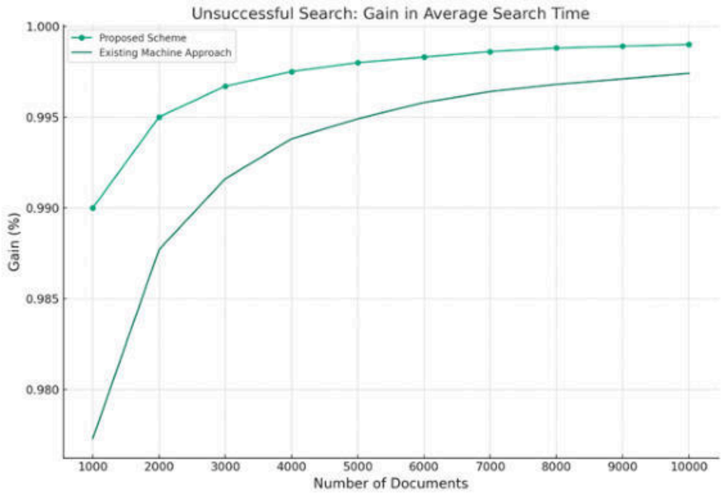


Fig. 5.6: Unsuccelful search: gain in Average search time

comparisons. The accompanying figure 5.6 illustrates the reduction in both the average number of comparisons and the average time required to determine an unsuccessful search. Compared to the previous schemes, the proposed method enhances the efficiency of declaring an unsuccessful search by 99.29%."

5.2. Computation Cost. In the proposed search scheme, the time required to construct the searchable index encompasses the duration to create both the text and the cluster indexes. The index build time for the proposed scheme tends to be greater than that of the existing scheme [33,42], which is attributed to the additional step of generating indexes for multiple clusters. The number of clusters, denoted by K, varies based on the application's needs, the size of the text collection, and the clustering algorithm employed. The incorporation of clustering into the index building process results in a slight increase in the time needed to create an index for a large text collection." For the graph, we assume we are illustrating the comparison of index build times between the proposed and existing schemes, highlighting the minor increase due to clustering. If you can provide any specific data or parameters you would like to include in the graph, please do so. Otherwise, we will proceed with a hypothetical representation. The average time required to construct a query, which encompasses HMAC computation, reduction, and bitwise-AND operations for each term, remains identical for both the proposed and the existing scheme [33,42], as the proposed scheme introduces no additional delays due to clustering. The average time to build queries, ranging from 1 to 5 genuine terms. , considering scenarios with and without the inclusion of noise terms. This timing is based on the mean time to generate 200 queries with a varying count of 1 to 5 genuine terms. To create a graph without specific values, we would plot the average query build time as it varies with the number of genuine terms for both scenarios (with and without noise terms). Since we don't have the specific values, we can create a hypothetical graph to illustrate this concept. Let's proceed to generate a graph that could represent this scenario see in the figure 5.7.

5.3. Rank Efficiency. The efficiency of result ranking is evaluated by comparing the time needed to generate per-text 'p' indexes at various relevance levels within the text collection. The increase in index build time associated with higher relevance levels is a one-time overhead, mitigated by the one-off nature of the indexing process conducted by the Data Owner (DO) during the offline stage. Cloud resources and parallel processing can be leveraged to further reduce this impact. Consequently, the extra time required for creating multiple indexes for each text is outweighed by the advantage of delivering superior ranked search results to the users. To visualize this concept, we can create a graph that demonstrates the efficiency of ranked search results without specifying exact values. Let's plot a graph showing the proportion of top-ranked texts from the proposed scheme that align with the top results from plain-text searches. We'll use hypothetical data to illustrate the concept described in figure 5.8.

6. Conclusion. The research conducted on the advancement in the field of secure recovery of data from cloud platforms. The proposed methodology integrates the resilience of greedy depth-first search algorithms with a sophisticated synonym identification system in order to offer accurate and efficient search functionalities across encrypted datasets. The method being described effectively addresses the challenges posed by synonymy and polysemy in search

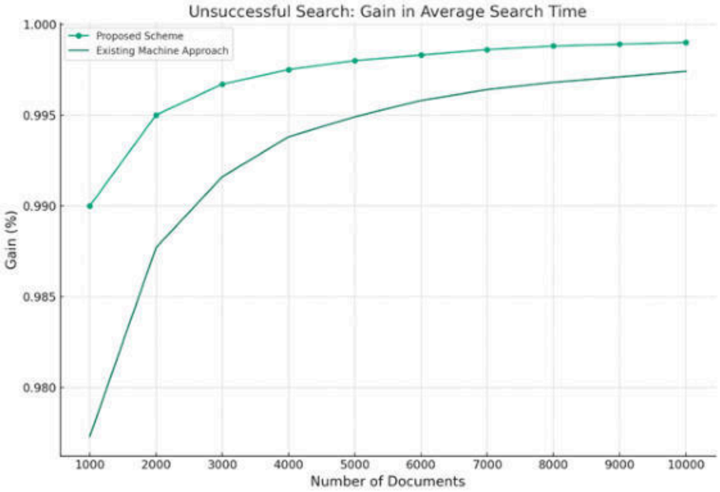


Fig. 5.7: Average query time by number of genuine term

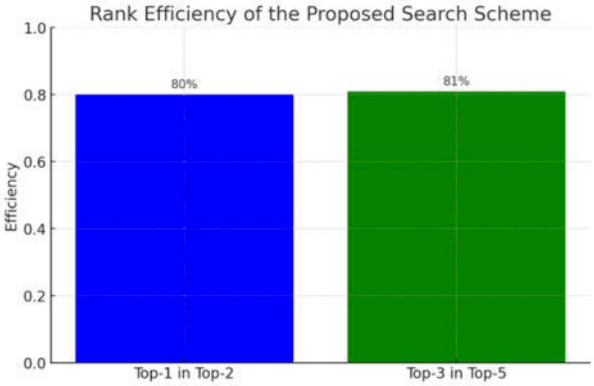


Fig. 5.8: rank efficiency of proposed search scheme

queries, thereby guaranteeing users access to comprehensive results that are not only pertinent to the specific terms employed but also to their semantic counterparts. The significance of this matter is particularly pronounced inside the realm of encrypted data, as conventional search methods are inadequate in light of the limits imposed by privacy preservation. the incorporation of a rating system inside the search process facilitates users in efficiently identifying the most relevant texts, hence augmenting the usability of cloud storage services. By implementing encryption techniques to handle privacy issues, while also ensuring a high degree of search accuracy and efficiency, the proposed method effectively fills a significant void in the utilisation of cloud data. The algorithm's efficacy, as evidenced by many performance measures, underscores its potential for extensive implementation in secure cloud-based applications. Given the escalating prevalence of cloud services, the concurrent rise in data privacy issues necessitates timely and crucial study to safeguard data security and accessibility in the future. Future research endeavours may further enhance this groundwork by delving into machine learning algorithms to achieve more refined synonym detection. Additionally, adaptive ranking techniques based on user feedback might be explored to optimise the system's performance. Furthermore, the scalability of the system should be investigated in light of the escalating demands for cloud storage. The continuous endeavour to achieve perfection in the development of search systems that are secure, efficient, and intelligent poses a persistent challenge. This research serves as a significant advancement in this ongoing goal.

REFERENCES

[1] P. Balamurugan, S. T. M, G. Arulkumaran and S. Jayagopalan, Multi-Keyword Graded Exploration in Encrypted Cloud Data for Industries Based on Rc4+ and Forest, *2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP), BHOPAL, India*, 2023, pp. 531-535, Doi: 10.1109/IHCSP56702.2023.10127116.

[2] C. Chi, Z. Yin, Y. Liu and S. Chai A Trusted Cloud–Edge Decision Architecture Based on Blockchain and MLP for AIoT, *IEEE Internet of Things Journal*, Vol. 11, no. 1, pp. 201-216, 2024, Doi:10.1109/JIOT.2023.3300845.

[3] L. Chen, Z. Chen, K. -K. R. Choo, C. -C. Chang and H. -M. Sun, Memory Leakage-Resilient Dynamic and Verifiable Multi-Keyword Ranked Search on Encrypted Smart Body Sensor Network Data, *IEEE Sensors Journal*, Vol. 19, no. 19, pp. 8468-8478, 1 Oct.1, 2019, Doi: 10.1109/JSEN.2018.2865550.

[4] H. Cui and X. Yi Secure Internet of Things in Cloud Computing via Puncturable Attribute-Based Encryption With User Revocation, *IEEE Internet of Things Journal*, Vol. 11, no. 2, pp. 3662-3670, 2024, Doi: 10.1109/JIOT.2023.3297997.

[5] D. Das, R. Amin and S. Kalra Algorithm for Multi Keyword Search Over Encrypted Data in Cloud Environment, *2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 2020*, pp. 733-739, Doi:10.1109/IWCMC48107.2020.9148472.

[6] D. Das and S. Kalra An Efficient LSI Based Multi-keyword Ranked Search Algorithm on Encrypted Data in Cloud Environment, *2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus*, 2020, pp. 1777-1782, Doi: 10.1109/IWCMC48107.2020.9148123.

[7] W. Dai et al. PRBFPT: A Practical Redactable Blockchain Framework With a Public Trapdoor, *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp. 2425-2437, 2024, Doi: 10.1109/TIFS.2024.3349855.

[8] Z. Fu, F. Huang, K. Ren, J. Weng and C. Wang, Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data, *IEEE Transactions on Information Forensics and Security*, Vol. 12, no. 8, pp. 1874-1884, Aug. 2017, Doi: 10.1109/TIFS.2017.2692728.

[9] Z. Fu, X. Wu, C. Guan, X. Sun and K. Ren, Toward Efficient Multi-Keyword Fuzzy Search Over Encrypted Outsourced Data With Accuracy Improvement, *IEEE Transactions on Information Forensics and Security*, Vol. 11, no. 12, pp. 2706-2716, Dec. 2016, Doi: 10.1109/TIFS.2016.2596138.

[10] Z. Fu, K. Ren, J. Shu, X. Sun and F. Huang, Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, no. 9, pp. 2546-2559, 1 Sept. 2016, Doi: 10.1109/TPDS.2015.2506573.

[11] Z. Fu, K. Ren, J. Shu, X. Sun and F. Huang, Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, no. 9, pp. 2546-2559, 1 Sept. 2016, Doi:10.1109/TPDS.2015.2506573.

[12] Z. Fu, X. Sun, N. Linge and L. Zhou, Achieving effective cloud search services: multi-keyword ranked search over encrypted cloud data supporting synonym query, *IEEE Transactions on Consumer Electronics*, Vol. 60, no. 1, pp. 164-172, February 2014, Doi:10.1109/TCE.2014.6780939.

[13] J. Han, L. Qi and J. Zhuang Vector Sum Range Decision for Verifiable Multiuser Fuzzy Keyword Search in Cloud-Assisted IoT, *EEE Internet of Things Journal*, Vol. 11, no. 1, pp. 931-943, 2024, Doi:10.1109/JIOT.2023.3288276.

[14] H. He, J. Liu, J. Gu and F. Gao An Efficient Multi-Keyword Search Scheme over Encrypted Data in Multi-Cloud Environment, *2022 IEEE 7th International Conference on Smart Cloud (SmartCloud), Shanghai, China*, 2022, pp. 59-67, Doi: 10.1109/SmartCloud55982.2022.00016.

[15] C. Huang, D. Liu, A. Yang, R. Lu and X. Shen Multi-Client Secure and Efficient DPF-Based Keyword Search for Cloud Storage, *IEEE Transactions on Dependable and Secure Computing*, Vol. 21, no. 1, pp. 353-371, 2024, Doi: 10.1109/TDSC.2023.3253786.

[16] A. Hosseingholizadeh, F. Rahmati, M. Ali, H. Damadi and X. Liu Privacy-Preserving Joint Data and Function Homomorphic Encryption for Cloud Software Services *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 728-741, 2024, Doi: 10.1109/JIOT.2023.3286508.

[17] B. Lang, J. Wang, M. Li and Y. Liu, Semantic-based Compound Keyword Search over Encrypted Cloud Data, *IEEE Transactions on Services Computing*, Vol. 14, no. 3, pp. 850-863, 2021, Doi:10.1109/TSC.2018.2847318.

[18] J. Li, J. Ma, Y. Miao, R. Yang, X. Liu and K. -K. R. Choo Practical Multi-Keyword Ranked Search With Access Control Over Encrypted Cloud Data, *IEEE Transactions on Cloud Computing*, Vol. 10, no. 3, pp. 2005-2019, 2022, Doi: 10.1109/TCC.2020.3024226.

[19] X. Liu, G. Yang, W. Susilo, J. Tonien, X. Liu and J. Shen, Privacy-Preserving Multi-Keyword Searchable Encryption for Distributed Systems, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 32, no. 3, pp. 561-574, 1 March 2021, Doi: 10.1109/TPDS.2020.3027003.

[20] Y. Miao, Y. Yang, X. Li, L. Wei, Z. Liu and R. H. Deng Efficient Privacy-Preserving Spatial Data Query in Cloud Computing, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, no. 1, pp. 122-136, 2024, Doi: 10.1109/TKDE.2023.3283020.

[21] B. M. Nguyen et al. A Novel Nature-Inspired Algorithm for Optimal Task Scheduling in Fog–Cloud Blockchain System, *IEEE Internet of Things Journal*, Vol. 11, no. 2, pp. 2043-2057, 2024, Doi:10.1109/JIOT.2023.3292872.

[22] P. Pandiaraja and P. Vijayakumar, Efficient Multi-keyword Search over Encrypted Data in Untrusted Cloud Environment, *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, India*, 2017, pp. 251-256, Doi:10.1109/ICRTCCM.2017.54.

[23] S. Prakash, N. Andola and S. Venkatesan, Secure access of multiple keywords over encrypted data in cloud environment using ECC-PKI and ECC ElGamal, *2017 International Conference on Public Key Infrastructure and its Applications (PKIA), Bangalore, India*, 2017, pp. 49-56, Doi:10.1109/PKIA.2017.8278960.

[24] V. Saiharitha and S. J. Saritha, A privacy and dynamic multi-keyword ranked search scheme over cloud data encrypted, *2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India*, 2016, pp. 1-5, Doi: 10.1109/CESYS.2016.7890001.

[25] V. Saiharitha and S. J. Saritha, A privacy and dynamic multi-keyword ranked search scheme over cloud data encrypted, *2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India*, 2016, pp. 1-5, Doi:10.1109/CESYS.2016.7890001.

[26] S. Shete and N. Dongre, Ranked multi-keyword search data using cloud, *2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India*, 2017, pp. 590-594, Doi:10.1109/ICICI.2017.8365200.

[27] M. Song, Z. Hua, Y. Zheng, H. Huang and X. Jia LSDedup: Layered Secure Deduplication for Cloud Storage, *IEEE Transactions on Computers*, vol. 73, no. 2, pp. 422-435, 2024, Doi: 10.1109/TC.2023.3331953.

[28] N. Wang, W. Zhou, J. Wang, Y. Guo, J. Fu and J. Liu Secure and Efficient Similarity Retrieval in Cloud Computing Based on Homomorphic Encryption *IEEE Transactions on Information Forensics and Security*, Vol. 19, pp. 2454-2469, 2024, Doi: 10.1109/TIFS.2024.3350909.

[29] Z. Xia, Q. Gu, W. Zhou, L. Xiong, J. Weng and N. Xiong STR: Secure Computation on Additive Shares Using the Share-Transform-Reveal Strategy, *IEEE Transactions on Computers*, Vol. 73, no. 2, pp. 340-352, 2024, Doi:10.1109/TC.2021.3073171.

[30] Z. Xia, X. Wang, X. Sun and Q. Wang, A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, no. 2, pp. 340-352, 1 Feb. 2016, Doi: 10.1109/TPDS.2015.2401003.

[31] D. Xu, C. Peng, W. Wang, K. Dev, S. A. Khowaja and Y. Tian, Multi-keyword Ranked Search Scheme Supporting Extreme Environments for the Internet of Vehicles, *IEEE Internet of Things Journal*, Doi:10.1109/JIOT.2023.3275386.

[32] Z. Yang et al. Differentially Private Federated Tensor Completion for Cloud–Edge Collaborative AIoT Data Prediction, *IEEE Internet of Things Journal*, Vol. 11, no. 1, pp. 256-267, 2024, Doi:10.1109/JIOT.2023.3314460.

[33] X. Yang, G. Chen, M. Wang, T. Li and C. Wang, Multi-Keyword Certificateless Searchable Public Key Authenticated Encryption Scheme Based on Blockchain, *IEEE Access*, Vol. 8, pp. 158765-158777, 2020, Doi: 10.1109/ACCESS.2020.3020841.

[34] H. Yin et al., Secure Conjunctive Multi-Keyword Search for Multiple Data Owners in Cloud Computing, *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS), Wuhan, China*, 2016, pp. 761-768, Doi: 10.1109/ICPADS.2016.0104.

[35] H. Yin et al., Secure Conjunctive Multi-Keyword Search for Multiple Data Owners in Cloud Computing, *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS), Wuhan, China*, 2016, pp. 761-768, Doi:10.1109/ICPADS.2016.0104.

[36] Y. Zhang, C. Jiang and P. Zhang Security-Aware Resource Allocation Scheme Based on DRL in Cloud–Edge–Terminal Cooperative Vehicular Network, em IEEE Internet of Things Journal, Vol. 11, no. 1, pp. 95-104, 2024, Doi: 10.1109/JIOT.2023.3293497.

[37] X. Zhou, D. He, J. Ning, M. Luo and X. Huang Single-Server Public-Key Authenticated Encryption With Keyword Search and Its Application in IIoT, *IEEE Transactions on Network Science and Engineering*, Vol. 11, no. 1, pp. 404-415, 2024, Doi: 10.1109/TNSE.2023.3300716.

Edited by: SCPE Editor-in-chief
Regular paper
Received: Feb 3, 2024
Accepted: Dec 13, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 490–502, DOI 10.12694/scpe.v26i1.3955

ENERGY AND DEADLINE AWARE WORKFLOW SCHEDULING USING ADAPTIVE REMORA OPTIMIZATION IN CLOUD COMPUTING

VIDYA SRIVASTAVA, RAKESH KUMAR *

Abstract. Cloud computing has become a more popular and well-known computing paradigm for delivering services to different organizations. The main benefits of the cloud computing paradigm, including on-demand services, pay-as-per-use policy, rapid elasticity, and so on, make cloud computing a more emerging technology to lead with new methods. Cloud systems have become more challenging than other systems because of their wide range of clients and the variety of services in the system. The cloud data center consists of many physical machines (PM) with virtual machines (VM), load balancers, switches, storage etc. Because of the inappropriate use of resources and inefficient scheduling, these data centers consume a lot of energy. In this paper, a multi-objective optimization model called Adaptive Remora Optimization (AROA) is proposed, which comprises sub-models viz; priority calculation, task clustering, probability definition and task-VM mapping using search mode based on Remora optimization to optimize energy consumption and execution time. CloudSim is used for the implementation of the proposed optimization technique. Through simulation the energy consumption is 0.695kWh and the execution time is 179.14sec. The result obtained by AROA is compared with the existing approaches to prove the efficacy of the proposed approach.Experimental results show that the proposed AROA algorithm outperforms the existing approaches.

Key words: Cloud computing, Execution time, Energy consumption, Remora optimization, Task scheduling

1. Introduction. Cloud computing has become vital to IT-based organizations and individual users over the last few years [15]. It gives desired computing assets for multiple programs in virtual machines, software, and hardware delivered by cloud service via cloud data centers [14]. Growing demands on cloud resources are completed with more powerful servers and other related hardware assets. Cloud computing enables the delivery of on-demand services such as networking, software, and intelligence via the Internet [39]. SaaS (Software as a service), PaaS (Platform as a service), and IaaS (Infrastructure as a service) are the three primary services used in the cloud computing environment, which provide extensible payment services, public-private, hybrid, and community cloud are four deployment models . Properties having multilevel abstraction and virtualization make cloud more resourceful computing [16]. The user only focuses on the utilization of resources and does not bother about their physical location. To improve the efficacy of the task scheduling algorithm user performed the proper allocation of a task to an appropriate resource. Scheduling plays a significant role in cloud computing [11]. Scheduling can be referred as an essential operating system function. All the resources that are available for operation should be scheduled before operation. Task scheduling is responsible for giving access to the system resources with the help of thread and process. So, task scheduling problems can be considered the appropriate obtaining mapping between the set of tasks over the available resources (CPU, bandwidth, virtual machine) [9, 38] . There are three types of task scheduling algorithms: pre-emptive, non-pre-emptive, and round-robin scheduling. When a high-priority task arises, a pre-emptive task scheduling algorithm interrupts the operation [32]. In non-pre-emptive task scheduling running task is executed till its completion. The third one is the round-robin scheduling algorithm that follows the FCFS policy in which pre-emption occurs in the middle of the operation [28]. In a scheduling-based scheme, all the tasks and jobs are appropriately arranged without interruption. Work must be completed within the deadline and executed individually to maximize reliability and minimize the execution time to optimize the system's overall time for allocating resources [30, 16].

Task scheduling plays a significant role in the cloud computing system. It cannot be done based on one criterion but on the various rules and regulations that can be termed as an accord between the user and cloud provider. The most crucial method to tackle the challenge of reducing energy consumption in the cloud environment can be done through effective task scheduling.[19]. According to the agreement, high-quality

*Madan Mohan Malaviya University of Technology, Gorakhpur, India (2020028010@mmmut.ac.in, rkiitr@gmail.com)

services to users and clients are the determining task for providers. At the same time, numerous jobs are running at the provider. Centralized and distributed are two main types of cloud task scheduling[31]. In a centralized task scheduling scenario, a single scheduler assigns tasks to resources, but if the scheduler fails, all systems are shut down. Additionally, in a centralized task scheduling system, scalability problems and fault tolerance issues occur[18]. In distributed task scheduling algorithms, schedulers are connected, meaning that all schedulers are collectively assigned to the tasks [6]. The single point of failure problem has been solved in distributed scheduling, but another complication, i.e., congestion problem, occurred between schedulers [1].

1.1. Motivation and Contribution. One of the significant awareness gaps associated with the previous survey is mapping tasks with appropriate VM placement in cloud computing. Several challenges arise in delivering services to the cloud user. There is a need for an efficient scheduling process provided by the cloud scheduler, such as SLA parameter, time, and cost must satisfy the user constraint[22]. In the case of a task, scheduling is concerned with the optimal mapping of resources on the specified virtual machine so that better system performance must be achieved. Variations in task characteristics and resource heterogeneity scheduling lead to NP-complete problems. No method is specified for finding the polynomial solution for such a problem. The scheduling algorithm reduced the makespan of the application[2]. Workflow scheduling is considered more with the emerging growth of cloud computing technology. A scheduling algorithm minimizes the makespan of the data center. However, sometimes selected energy does not give the optimal solution, so task scheduling can be treated as the more challenging way to increase the reliability of a system and reduce the makespan simultaneously [4]. To overcome this problem, an optimization function/ fitness function is required for finding the suitable scheduling. In the cloud environment, scheduling issues become widely explored, whether it is to be single or multi objectives challenges [37]. A large no. of applications with their higher load makes the cloud more complex due to the inefficient use of resources. It is a challenge for researchers to present a new optimization approach in a dynamic environment that overcomes the problem in previous studies. The user aims to find suitable physical hosts for their users, and cloud providers aim to utilize their infrastructure.

2. Related Works. The author of [26] proposed a workflow scheduling system that is both an energy-efficient and reliable(EERS) mechanism that jointly optimizes the system's dependability and minimizes energy consumption. EERS consists of five sub-algorithms; first, a rank calculation algorithm is used to preserve the dependency of a task. Second is the clustering algorithm for conserving energy. The third discusses a distribution mechanism for defining the makespan for each task, then applies the fourth method i.e., cluster-VM mapping aiming to maximize the system's reliability and minimize energy consumption. The last algorithm is a slack algorithm associated with non-critical tasks. The simulation result indicates that the proposed algorithm optimizes energy and reliability in polynomial time. It also examined the genuine electricity cost for the task scheduling algorithm as a subsequent work.

The paper [12] developed a new scheduling approach known as reliability and energy-efficient workflow scheduling (REEWS) to maximize a system's reliability and minimize energy consumption. REEWS Scheme is divided into four sub-algorithms. The first is priority calculation, the second is the clustering of the task, the third is distribution, and the fourth is assigning clusters with a proper frequency level. Gaussian elimination and a randomly generated graph are used to enhance an algorithm's performance. The performance of the REEWS algorithm is compared with other well-known such as heterogeneous Earliest-finish-time(HEFT), reliable-HEFT(RHEFT), algorithms. It will have observed that the REEWS algorithm has outstanding performance compared to the different algorithm simulated result will be enhanced by combining the clustering algorithm with the load balancing module.

The survey [24] proposed an in-depth analysis of PSO scheduling, main objectives such as load balancing, makespan, and execution time. Particle swarm optimization is a population-based meta-heuristic technique covering a wide range of applications because of its effectiveness and low computational cost. In addition to that, different levels of trust and reliability must be investigated and evaluated to solve more scheduling problems.

The author of [25] developed a task scheduling optimization scheme based on improved ant colony optimization to improve scheduling methods that fall under local optimization. The fitness function determines the optimal solution for the scheduling algorithm. The feasibility analysis demonstrates how the algorithm performs well with the fastest convergence speed and shortest computation time.

The paper [5] introduced a technique based on the co-optimization process aiming to map the task into a virtual machine within the deadline time and then assign a suitable placing virtual machine to the correct physical host within the capacity constraint. Experimental results show how this technique performs better than all other optimization problems.

Another paper [29] proposed algorithm works in two stages: Virtual machine(VM) scheduling and consolidation. The maximum runtime job is assigned to the virtual machine during the VM scheduling phase, which is expected to reduce energy usage. A double threshold technique is used in the consolidation phase to find overload and underload hosts. Experimental results simulate the Combination of scheduling and consolidation phase successfully increases resource utilization and decreases energy consumption.

The study done by [35] discussed a comprehensive review of meta-heuristic optimization algorithms in the cloud computing system. Approaches discussed in the metaheuristic mechanism will be enough for the reader to select a new mechanism for solving task scheduling problems. A brief review of future research work was also discussed in the research work.

The author of [36] presents two heuristic algorithms, including budget-deadline constraints. Resources that support DVFS technology use Budget-deadline DVFS enable energy (BDD); resources that do not work well with DVFS use Budget-deadline constraint energy aware (BDCE). Several metrics like cost, utilization rate, energy consumption, and success rate are utilized to estimate the fulfillment of the proposed algorithm.

The paper [13] presents a comprehensive review of energy optimization in a cloud computing environment that compares 67 scheduling algorithms that reduce energy consumption throughout the scheduling process. This work is appropriate for the reader to select a relevant approach to minimizing energy consumption.

The study done in [3] developed a multi-workflow scheduling algorithm with dynamic reusability aiming to minimize the time taken to complete individual tasks and then computing the overall time taken to complete tasks within the deadline; they dynamically reuse the available virtual machine while it is needed Because of the maximum utilization of resources such as CPU and virtual machines within their deadline, simulation results show that the algorithm is moving towards a new generation of multi-objective scheduling. Table 2.1 shows the Silent features of a few existing task-scheduling algorithms.

3. System Model.

This section introduced various types of system models, including the cloud model, task model, workflow model, and energy model detailed illustration has been done in the proposed work. The Cloud sim toolkit is used for simulation to implement the proposed work. For easy insight, Table 3.1 outlines the primary notations with their meaning, which are used in this research paper.

3.1. Cloud Model.

In general cloud system can be represented by infrastructure as a service (IaaS) which is responsible for managing resources , such as physical and virtual, to fulfilling the requirement of cloud users. Researchers were aiming to build a system model to enhance Various parameters available to measure the capacity of a system, such as CPU, RAM, Bandwidth, and storage [34, 23]. For scheduling, the execution of workflow virtual machine is assigned to each physical host[32]. A physical host can carry more than one virtual machine based on requirements. We consider M number of virtual machines available that can be represented as VM= VM_1, VM_2, VM_3.... VM_v, which are fully connected with each other. Fig. 3.1 shows the cloud architecture.

3.1.1. Task Model.

Task scheduling is defined as assigning a task to the appropriate resources to minimize energy consumption and execution time. Several models are available depending on the scheduling criteria available on the specific cloud system. Consider m number of resources present in cloud represented as R= r_1,r_2,r_3....r_c, and mapping of task is denoted by M f : $T \rightarrow R$ represent the mapping function such as Mf (j) ,represent resources corresponding to the task t_k is assigned[10]. Fig. 3.2 shows the task model The virtual machine is the collection of computing resources that helps to virtualize the physical machine.

3.1.2. Energy Model.

In the data center model, resources such as CPU and other networking devices can cause energy consumption. From the net amount of energy, only processors have depleted 37-43% In the data center model, resources such as CPU and other networking devices can cause energy consumption. From the net amount of energy, only processors have depleted 37-43% Networking devices consumed 33% of total energy[33]. As we all know, all networking devices are nearly fixed and cannot be modified in the event of any execution workflow. When the processor begins to function, the energy consumption is directly proportional

Table 2.1: Silent features of existing task scheduling algorithms

Authors and year of Publication	Key objective	Application area	Issues
Ref.[[26]](2021)	Minimizing energy consumption maximizes system reliability.	EERS algorithms have been proposed for workflow scheduling and consolidation.	Cost,frequency independent energy consumption
Ref.[[12]](2019)	Perform a combinatorial optimization of the reliability of the application and guarantees the QoS.	Proposed a REEWS Algorithm executes in four stages: calculation, clustering, time distribution associated with assigning appropriate voltage	Only consider essential features of scheduling algorithms.
Ref.[[5]] (2021)	Minimizing execution cost, degree of Imbalance, makespan, and cost.	Aims to assign scheduled tasks into appropriate VM with the least execution cost within deadline constraints.	More memory required to increase complexity.
Ref.[[24]](2017)	Minimize energy consumption and achieve green cloud computing .	Presents a deep analysis of the particle swarm optimization algorithm.	Not consider Load balancing and VM parameters.
Ref.[[37]](2020)	Increase convergence speed, and minimize completion time,	Propose an improved ant colony algorithm, including pheromone update and volatilization	Resource utilization, ensuring QoS and energy consumption.
Ref.[[29]](2018),	Maximize resource utilization, minimize energy consumption.	The migration technique is used for the detection of overload and underload hosts.	Unable to find a better resource selection.
Ref.[[35](2017)]	Suggest readers decide suitable approach for scheduling	Present a competitive taxonomy among existing algorithms.	Lack of scalability and standardization in new computing model.
Ref. [[36]](2021)	Affordable price increasesscheduling length and higher energy-saving ratio.	Develop two heuristic algorithms suitable for cloud environments, including budget and deadline constraints.	High-fault occurrence, high resource wastage.
Ref.[[13]](2022)	Comparative analysis of 67 algorithms to minimize energy consumption.	Discuss important aspects of heuristics, metaheuristics, and task scheduling algorithms.	Need for a simulator that supports multi-objective scheduling.
Ref. [[3](2018)	Reduce the makespan of every task and wrap up the overall execution of the entire workflow within the deadline.	Present competent and dynamic multi-workflow scheduling (CDMWS) algorithm that dynamically produces VM.	Increases network load, rescheduling task

to the execution of the workflow. When the resources are equipped, static and dynamic energy are responsible for energy consumption; static energy remains constant, whereas dynamic energy is entirely dependent upon the frequency/voltage of the corresponding processor. In the case of frequency transition, we consider zero overhead because each transition takes a minute amount of time (about the microsecond range). It is more precise to say that dynamic energy is blamed for energy consumption.

3.1.3. Workflow Model. Workflow can be defined as the group of computational tasks with their reliance constraint between them[20]. Data passing from one workflow to another is considered a dependency constraint. So, DAG D = (T, E) where T= collection of task (t_1, t_2,t_3.....t_k) and E = collection of edge(e_1,e_2,e_3,.....e_g,) which indicate the correlation between the task. The graph is used to dictate the workflow task. It is an automation of a repeatable pattern of processes where the data and information are passed from one cloud user to another for specific action. To improve efficiency and profitability, there is a need for coordination between

Table 3.1: Primary Notations

Notation	**Meaning**
EnC	Energy Consumption
ExT	Execution time
$EnC_dynamic$	Dynamic energy consumption
$EnC_Sstatic$	Static energy consumption
t_k	The kth task in workflow
r_c	The Cth resource in workflow
Cc	Computing capacity
Vol	Voltage
f	Frequency
M	Total number of virtual machines
Cp	Capacitance
P	Remora Position in search space
Dist	Current optimal solution
n	Represents the number of remora
i	Current iteration
I	Maximum iteration
Pbest	The optimal solution in the algorithm
f(Pbest)	The fitness function of the best position
P_n	Current position
P_{pre}	Position of the previous iteration
P_{tnt}	Tentative step
P_{random}	Random location
randn	Small global movement
A	Volume space

the user and synchronized data. In workflow execution criteria, workflow scheduling plays a major key issue; however, workflow defines the execution of workflows on which tasks are assigned to well-suited resources for satisfying constraints such as energy consumption and execution time[27].

3.2. Problem Formulation. We consider independent scheduling tasks that comprise heterogeneous virtual machines (VM) and physical machines (PM). This section introduces an optimization function that consists of reducing energy consumption as well as execution time and also represents constraints that are specified in this problem. The main aim is to find a suitable algorithm for workflow scheduling and virtual machine placement to reduce energy consumption and execution time.

3.2.1. Energy Consumption. In this study, we opted classic energy consumption model to analyze the energy consumption.

$$EnC = EnC_{static} + EnC_{dynamic} \tag{3.1}$$

where EnCstatic is energy consumption when the system does not carry out any workload, i.e., the system turned off. We consider the dynamic consumption of the cloud model to be discussed in the model. Total Energy consumption is defined in Eq. (3.1). Dynamic energy consumption occurs when the system turns on. Formulations Cp, Vol and f, represent constants belonging to processor capacity, voltage, and frequency, respectively, defined in Eq. (3.2).

$$EnC_{dynamic} = Cp.Vol^2.f \tag{3.2}$$

where C, V and f represent capacity voltage and frequency, respectively. Since $f \propto vol^{\frac{1}{\eta}} for(0 < \eta < 1)$ or i.e frequency-dependent energy consumption defined in Eq.(3.3)

$$EnC_{dynamic} \propto f^{\lambda} \quad \lambda = 1 + 2/\eta \geqslant 3 \tag{3.3}$$

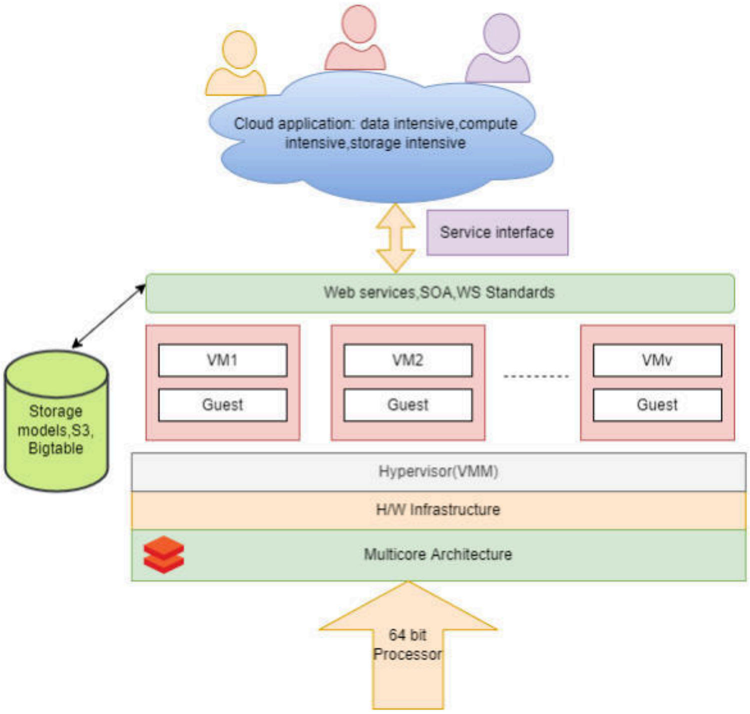

Cloud application: data intensive,compute intensive,storage intensive
Service interface
Web services,SOA,WS Standards
Storage models,S3, Bigtable
VM1
Guest
VM2
Guest
VMv
Guest
Hypervisor(VMM)
H/W Infrastructure
Multicore Architecture
64 bit Processor

Fig. 3.1: Cloud Architecture

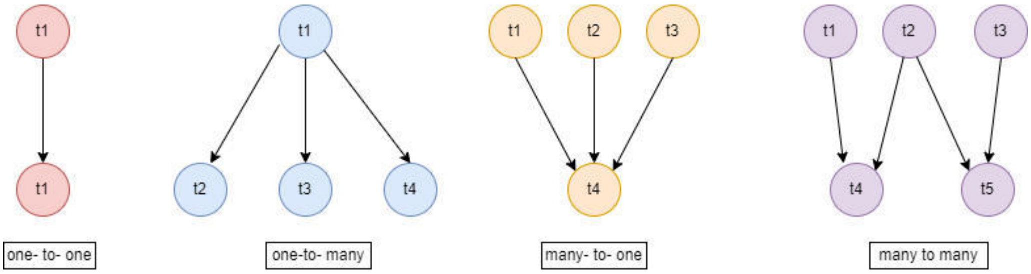


Fig. 3.2: Task Model

So, in this study, energy consumption is represented in Eq. (3.4)

$$EnC = EnC_{static} + Cp.f^{\lambda} \tag{3.4}$$

Further, Energy consumption can be divided into two parts-resource is busy/idle. Processor frequency will be at min level when the resources are idle. Processor frequency will be highest in case of execution starts, defined in Eq. (3.5)

$$EnC_{r_c}[total] = EnC_{r_c}[idle] + EnC_{r_c}[busy] \tag{3.5}$$

when resource rc is busy (i.e, the task is executing), energy consumption is calculated by Eq. (3.6), where Vol and f represent the voltage and frequency of resources rc upon the task tk is being finished their execution.

$$EnC_{r_c}[busy] = \sum_{t_k \in r_c} C_p.Vol^2_{r_c,t_k}.f_{r_c,t_k}.E_xT \tag{3.6}$$

When a resource is idle (i.e., the task is not executing), the processor works at a minimum frequency and voltage. During an idle time of resource, energy consumption is defined in equation (3.7), where idlerm defined idle time of rm.

$$EnC_{r_c}[idle] = C_p.Vol^2_{min}.f_{min}.idle_{r_c} \tag{3.7}$$

Therefore total energy consumption can be calculated in equation(3.8).

$$EnC_{r_c}[total] = \sum_{t_k \in r_c} Cp.Vol^2_{r_c,t_k}.f_{r_c,t_k}.E_xT + Cp.Vol^2_{min}.f_{min}.idle_{r_c} \tag{3.8}$$

3.2.2. Execution Time. The Execution time ExT denotes the time interval that is taken for executing task tk on resources VMCp based on the computing capacity of available resources that is defined as in Eq.(3.9):

$$ExT(t_k, VM_{Cc}(t_k)) = \frac{\sum_{i=0}^{k} length(t_k)}{\delta(VM_{Cc})} \tag{3.9}$$

3.3. Proposed Mechanism. We proposed a new scheduling algorithm called Adaptive Remora Optimization Algorithm(AROA) to minimize energy consumption and execution time (Figure 3.3). The mechanism is explained in four phases:

1. Priority calculation to provide a reliable topological workflow ordering that satisfies precedence constraint.
2. Tasks clustering for reducing the communication cost among tasks aiming to minimize energy consumption of the given system.
3. Define a probability definition.
4. Assigning the tasks to the appropriate processor at the proper voltage/frequency level to minimize energy consumption and execution time. Fig.3.3 shows proposed mechanism.

3.3.1. Priority Calculation. The task priority is calculated to ensure that the most time-consuming tasks are finished first. The task priority order is calculated to ensure that the most tedious tasks are executed first. In addition to that, it saves tasks that are waiting for input from higher-priority tasks. The tasks are prioritized in a hierarchical order for scheduling. The tasks are scheduled so that the precedence constraints are met.

Algorithm 1 Priority Calculation for Task (t_k)

```
Initialize the number of tasks ($t_k$)
for each task do
    Evaluate execution time (ExT)
    Sort the execution time in ascending order
    Assign the priority such that the lowest execution time task gets the highest priority
end for
```

3.3.2. Clustering of Task. After priority calculation of all tasks, clustering will be the next step in which a cluster formation occurs, which will be used in remora scheduling.

3.3.3. Probability Definition. The proposed algorithm reduces the execution time by optimizing the available number of processors and energy consumed by the processor. The scheduling method divided into the no. of equal time steps (Δt_s). The probability-based scheduling algorithm determines the probability of execution time P(ExT) of the task into equal time steps (Δt_s). The task is scheduled at which time when the probability of execution time is minimum. The probability of execution time is the sum of its current task and its successor's tasks. It is assumed that all tasks are operated at a higher frequency[21]. The probability of execution time is dynamically updated based on time steps assigned to the selected voltage-scaled job. Finally, the no. of processors required to execute the last schedule is calculated. The steps are explained in the following:

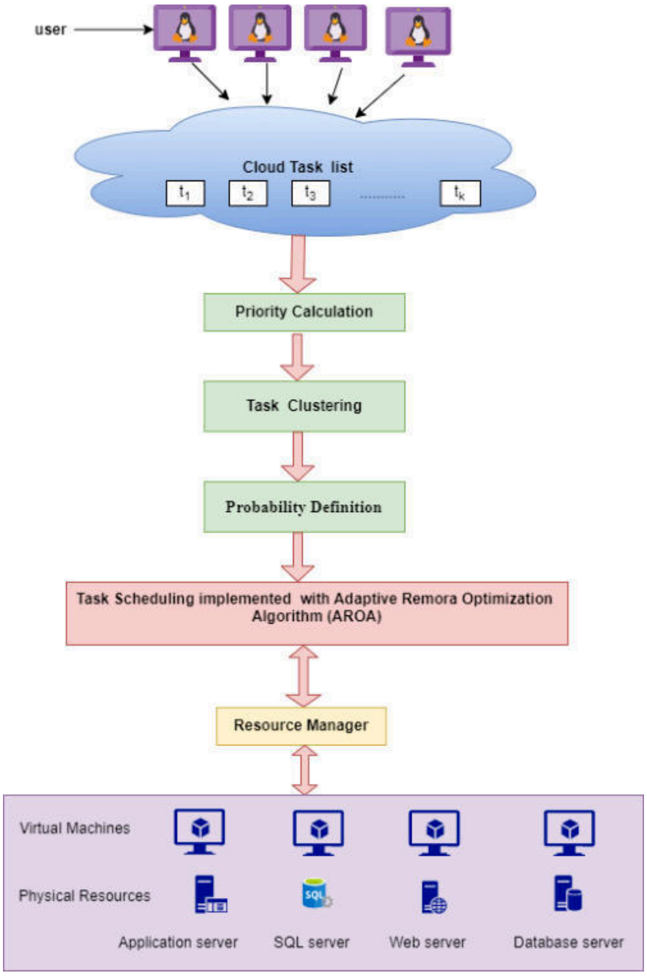


Fig. 3.3: Proposed Mechanism

Algorithm 2 K-Means Clustering Algorithm

```
Select the number of clusters, $k$
Select random tasks as initial centroids (other than the input dataset)
repeat
    Allocate each task to its closest centroid, forming $k$ clusters
    Evaluate variance and update centroids for each cluster
    Repeat the allocation and centroid update for more cluster formation
until No reassignment occurs
Cluster formation is complete
```

Step 1 (Time step calculation) Available tasks in cluster (t_1, t_2, t_3, t_4 ... t_k) is to be considered they are operated at the BCET scheduling algorithm. All the tasks are operated at a higher frequency. Tasks are divided into the time step taking as a unit with the smallest time step execution.

Step 2 (Best-case Performance) Using BFS (breadth-first search), the best-case algorithm begins with the source task in the task graph before it. In case of scheduling the predecessor, each task is scheduled with the earliest available time step BCP (t_k). The task is categorized by its start time BCPstrt(t_k).

and finish time BCPfn(t_k). The earliest possible execution time of the task graph is described in this algorithm.

Step 3 (Worst-case Performance) Worst-case scheduling same as the best-case scheduling algorithm. The difference is only its start with the sink of the task graph with upward preceding. Algorithms describe the longest possible execution of the task graph. Task(t_k) is categorized by its start time WCPstrt(t_k). and finish time WCPfn(t_k). This algorithm describes the latest possible execution time of the task graph.

Step 4 (Bound limit demonstration) Algorithm moves towards determining the bound limit (Δtk) of each task. The bound limit can be calculated in Eq. (3.10).

$$\Delta t_k = worstcase_{fn}(t_k) - bestcase_{strt}(t_k) \tag{3.10}$$

where Δt_k is the time step within which the current task is scheduled.

Step 5 (Finding the probability of execution time of each task in each time step) Next step is to find the probability of execution time in each time

3.4. Adaptive Remora Optimization Algorithm. ROA is a new natural bio-inspired and metaheuristic algorithm aiming to minimise energy consumption and execution time. Parasitic behavior is the main inspiration for the remora algorithm. Host locations are updated[17]. In the case of a large host (giant whales), remora feed on the host's extermination and natural enemies. In the case of a small host, the remora chase the host to move fast(swordfish) to the bait-rich area to prey; from the above two cases, the remora makes a judgment based on experience. When it starts initiation to prey, continuously update the host and make a global decision. If it eats encircling the host remora, continue to local update without changing their host. Fitness function f(x) is defined in Eq.(11)

A population-based search algorithm's main attainment is exploration and exploitation trade-off. AROA has two parts: exploration and exploitation like ROA. These two parts have a great impact on the algorithm that how long these two algorithms perform. The main difference between remora and adaptive remora optimization is utilizing a new parameter, "search mode(SM)." As mentioned in the AROA algorithm, the value decreases over time with a small value. In the initial search, space exploration changes the solution, activating the exploitation part on time increment. However, in AROA, the active part, including viz, exploitation and exploration, is altered adaptively by introducing the search mode parameter for tracking the behaviour of the solution in the given population. Assuming the search mode value is set to 1 for promoting the exploitation stage, the calculated probability from Probability Definition Step 5, In such a situation exploitation stage become more active than exploration stage. If the solution quality does not change over the iteration search mode is updated 2. in such a scenario exploration stage becomes more active compared to exploitation. According to algorithm 3, search mode SM not only determines the active part adaptively at a time, also justifies the duration when the exploration/exploitation parts become active and can change the position of solution.

4. Simulation Results and Experiments. CloudSim tool is used for evaluating the proposed AROA optimization algorithm. It is broadly used to simulate the cloud systems methods such as virtual machines and data centers to support task scheduling policies such as task selection and virtual machine placement [8]. It is to be very clear that the proposed AROA works on the large-scale data center. At the host level total no. of a heterogeneous host is taken to be 800 capacity of RAM is 512MB, and the corresponding bandwidth is to be 1000 Mbps, CPU capacity is 50 MIPS, and storage is 100GB. At VM level total no. of XEN virtual machine is 1175, no, of tasks is set to be 3000, the memory is 1536 MB, the CPU capacity is 1000 MIPS, no. of CPU is 1, the bandwidth is 1000 Mbps, and the storage is 1000 GB. Various varieties of performance are evaluated to calculate the performance of energy consumption and execution time. Based on metrics proficiency, the proposed algorithm can be calculated. A comparison between the existing and proposed approaches is performed[7].

4.1. Evaluation of Energy Consumption. The experimental results of the proposed AROA are to be discussed in this section. Adaptive remora optimization's performance is comparable with the well-known existing approach to find the efficiency of the proposed method. Existing mechanisms are the Combination of metaheuristic approaches such as Genetic algorithm (GA), Particle-swarm optimization (PSO) algorithm, Minimum-Migration Time (MMT) policy, Random selection (RS) policy, Median Absolute Deviation (MAD)

Algorithm 3 Adaptive Remora Optimization Algorithm

```
Input: Application graph $G(T, E)$, number of physical machines/processors
Output: Energy-efficient workflow scheduling
Step 1 Input the task $t_k$ in $T$
Step 2 Calculate priority from Algorithm 1
Step 3 Assign priority such that the lowest execution time task gets the highest priority
Step 4 Make clusters using Algorithm 2
Step 5 Calculate the number of clusters
Step 6 For each cluster
Step 7 Select Remora population size ($N$) and Max number of iterations ($I$)
Step 8 Set the position of entire search agents $P_n$ ($n = 1, 2, 3, \ldots, N$)
Step 9 While ($i \leq I$)
Step 10 Evaluate the fitness of each remora
Step 11 Evaluate the best fitness and best position, $P_{\text{best}}$
Step 12 $Prob_s$ = calculated using Probability Definition in Step 5; $SM = 1$; $K_{\text{rand}} = \text{rndreal}(0, 1)$
Step 13 For $n$th remora
Step 14 if $H(n) = 0$
Step 15 if ($SM == 1$ and $K_{\text{rand}} \leq Prob_s$) or ($SM == 2$ and $K_{\text{rand}} > Prob_s$) // exploits mode
Step 16    ...
Step 17    End if
Step 18    Else if $H(n) = 1$
Step 19    if ($SM == 2$ and $K_{\text{rand}} \leq Prob_s$) or ($SM == 1$ and $K_{\text{rand}} > Prob_s$) // explore mode
Step 20    ...
Step 21    End if
Step 22    End if
Step 23    Generate tentative candidate position
Step 24    ...
Step 25    If $f(P_{\text{tnt}}) < f(P_n)$
Step 26       $P_n = P_{\text{tnt}}$
Step 27       $H(n) = \text{round(random)}$
Step 28    Else
Step 29       Update
Step 30    Endif
Step 31 End for
Step 32 $i = i + 1$
Step 33 End While
Step 34 Return the best fitness value $P_{\text{best}}$
```

and Interquartile Range (IQR) algorithms. The number of virtual employed for executing workflow is the important factor for calculating that is based on energy consumption. Virtual machines take place inside the physical machine. Thus, energy consumption is closely associated accord to the number of the current physical server. Energy consumption is calculated based on Eq. (3.8).

Fig. 4.1 compares the proposed AROA energy consumption(in kWh) with eight existing approaches. Compared to the existing approaches, the proposed AROA consumes minimum energy throughout the entire execution in the cloud computing system. The proposed approach is efficient for workflow based on the experimental results.

4.2. Evaluation of Execution Time. Fig. 4.2 compares the execution time (in sec) of the proposed AROA with eight existing approaches. The proposed approach consumed less time in the cloud computing system when compared to the existing approaches. From the experimental results, the proposed approach is efficient for workflow scheduling.

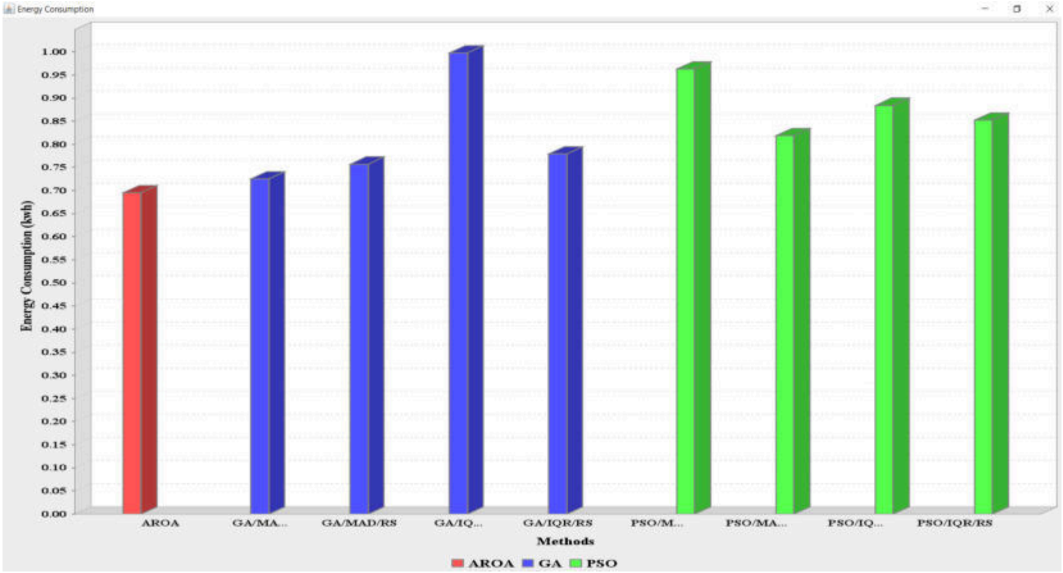


Fig. 4.1: Comparison of Proposed AROA EnC and eight existing approaches

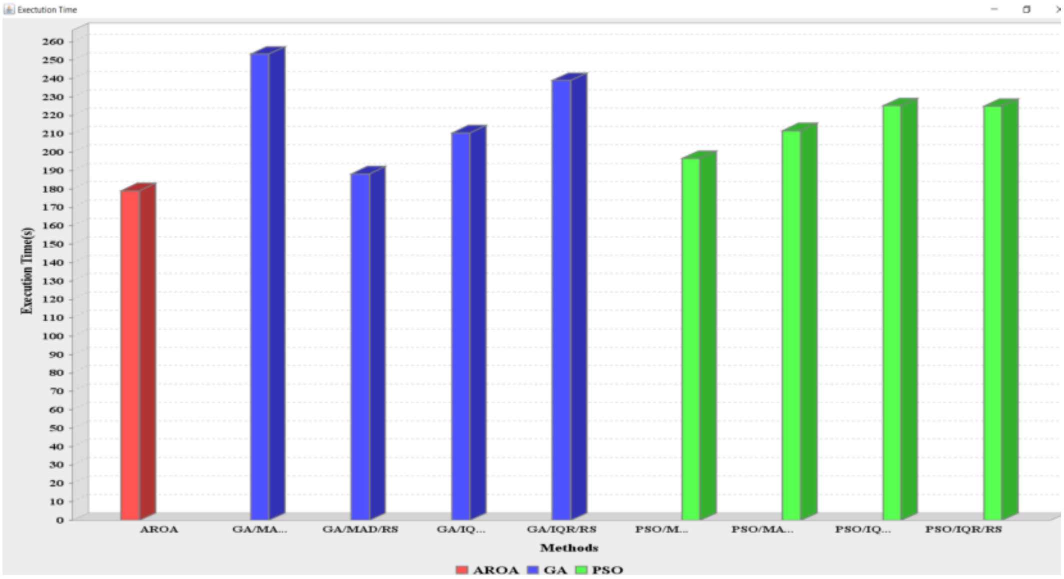


Fig. 4.2: Comparison of proposed AROA ExT and eight existing approaches

4.3. Result Discussion. The performance comparison of the proposed Adaptive remora optimization algorithm (AROA) was done versus the eight existing meta-heuristic algorithms, including GA/MAD/MMT, GA/MAD/RS, GA/IQR/MMT, GA/IQR/RS, PSO/MAD/MMT, PSO/MAD/RS, PSO/IQR/MMT, PSO/ IQR/RS. The performance analysis indicates that the proposed AROA provides better results when comparison is performed with the existing mechanism for task scheduling. Proposed AROA energy consumption is .695 kWh and execution time is 179.14sec.

4.4. General Computational Complexity Calculation. The computing complexity influences the initialization, fitness assessment, and position update technique of the optimization algorithm. Initialization has an O(N) computational complexity for the fundamental ROA. In this case, the number of search agents is represented by the parameter N. The computational complexity of applying the SFO or WOA method for the entire iterative process is O (N $\times$ D$\times$ I), where I is the maximum number of iterations and D is the dimensions of the search space.

5. Conclusion. In cloud computing, the workflow schedule is a significant process for assigning tasks to virtual machines. Energy consumption and execution time have recently become more important in research work. We Consider several tasks in a cloud environment. Here we present Adaptive Remora optimization (AROA), comprising sub-models viz; priority calculation, clustering, probability definition and task-VM mapping using search mode for optimizing energy consumption and execution time. The experimental result shows that adaptive remora is more suitable for minimizing energy consumption and execution time. CloudSim is the simulation platform for the use implementation of AROA optimization techniques. Performance of proposed algorithm compared with some existing approaches. The experimental result depicts , energy consumption is 0.695kWh, and the execution time is 179.14 sec. The result exhibits that the proposed technique is much better than the existing approaches. Future scope of improvement consists of

1. Load balance mechanism on VM.
2. Adding more data center on the network.
3. Including metrics such as cost-effectiveness, cloud mobility.
4. Proposed reliability and budget constraint scheduling algorithm.
5. Concepts of hybrid cloud and security factors should be considered more.

REFERENCES

[1] M. Abd Elaziz, S. Xiong, K. Jayasena, and L. Li, *Task scheduling in cloud computing based on hybrid moth search algorithm and differential evolution*, Knowledge-Based Systems, 169 (2019), pp. 39–52.

[2] M. N. Abdulredha, A. A. Bara'a, and A. J. Jabir, *Heuristic and meta-heuristic optimization models for task scheduling in cloud-fog systems: A review*, Iraqi Journal for Electrical And Electronic Engineering, 16 (2020), pp. 103–112.

[3] M. Adhikari and S. Koley, *Cloud computing: a multi-workflow scheduling algorithm with dynamic reusability*, Arabian Journal for Science and Engineering, 43 (2018), pp. 645–660.

[4] R. Al-Arasi and A. Saif, *Task scheduling in cloud computing based on metaheuristic techniques: A review paper*, EAI Endorsed Transactions on Cloud Systems, 6 (2020).

[5] D. Alboaneen, H. Tianfield, Y. Zhang, and B. Pranggono, *A metaheuristic method for joint task scheduling and virtual machine placement in cloud data centers*, Future Generation Computer Systems, 115 (2021), pp. 201–212.

[6] N. Anwar and H. Deng, *A hybrid metaheuristic for multi-objective scientific workflow scheduling in a cloud environment*, Applied sciences, 8 (2018), p. 538.

[7] M. Askarizade Haghighi, M. Maeen, and M. Haghparast, *An energy-efficient dynamic resource management approach based on clustering and meta-heuristic algorithms in cloud computing iaas platforms: Energy efficient dynamic cloud resource management*, Wireless Personal Communications, 104 (2019), pp. 1367–1391.

[8] J. Aswini, B. Yamini, R. Jatothu, K. S. Nayaki, and M. Nalini, *An efficient cloud-based healthcare services paradigm for chronic kidney disease prediction application using boosted support vector machine*, Concurrency and Computation: Practice and Experience, 34 (2022), p. e6722.

[9] A. Belgacem and K. Beghdad-Bey, *Multi-objective workflow scheduling in cloud computing: trade-off between makespan and cost*, Cluster Computing, 25 (2022), pp. 579–595.

[10] G. Bindu, K. Ramani, and C. S. Bindu, *Optimized resource scheduling using the meta heuristic algorithm in cloud computing*, IAENG International Journal of Computer Science, 47 (2020), pp. 360–366.

[11] R. Z. Frantz, S. Sawicki, F. Roos-Frantz, F. P. Basso, B. Zucoloto, and R. M. Pillat, *On the analysis of makespan and performance of the task-based execution model for enterprise application integration platforms: An empirical study*, Software: Practice and Experience, 52 (2022), pp. 1717–1735.

[12] R. Garg, M. Mittal, and L. H. Son, *Reliability and energy efficient workflow scheduling in cloud environment*, Cluster Computing, 22 (2019), pp. 1283–1297.

[13] R. Ghafari, F. H. Kabutarkhani, and N. Mansouri, *Task scheduling algorithms for energy optimization in cloud environment: a comprehensive review*, Cluster Computing, 25 (2022), pp. 1035–1093.

[14] T. K. Ghosh, K. G. Dhal, and S. Das, *Comparative study of some nature-inspired meta-heuristics for task scheduling in a computational grid system*, in Novel Research and Development Approaches in Heterogeneous Systems and Algorithms, IGI Global, 2023, pp. 1–15.

[15] I. M. Ibrahim et al., *Task scheduling algorithms in cloud computing: A review*, Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12 (2021), pp. 1041–1053.

[16] B. Jamil, H. Ijaz, M. Shojafar, K. Munir, and R. Buyya, *Resource allocation and task scheduling in fog computing and internet of everything environments: A taxonomy, review, and future directions*, ACM Computing Surveys (CSUR), 54 (2022), pp. 1–38.

[17] H. Jia, X. Peng, and C. Lang, *Remora optimization algorithm*, Expert Systems with Applications, 185 (2021), p. 115665.

[18] J. Kakkottakath Valappil Thekkepuryil, D. P. Suseelan, and P. M. Keerikkattil, *An effective meta-heuristic based multi-objective hybrid optimization method for workflow scheduling in cloud computing environment*, Cluster Computing, 24 (2021), pp. 2367–2384.

[19] S. Kaur, P. Bagga, R. Hans, and H. Kaur, *Quality of service (qos) aware workflow scheduling (wfs) in cloud computing: A systematic review*, Arabian Journal for Science and Engineering, 44 (2019), pp. 2867–2897.

[20] J. K. Konjaang and L. Xu, *Meta-heuristic approaches for effective scheduling in infrastructure as a service cloud: A systematic review*, Journal of Network and Systems Management, 29 (2021), pp. 1–57.

[21] A. A. Laghari, H. He, A. Khan, N. Kumar, and R. Kharel, *Quality of experience framework for cloud computing (qoc)*, IEEE Access, 6 (2018), pp. 64876–64890.

[22] A. A. Laghari, A. K. Jumani, and R. A. Laghari, *Review and state of art of fog computing*, Archives of Computational Methods in Engineering, (2021), pp. 1–13.

[23] N. Manikandan, N. Gobalakrishnan, and K. Pradeep, *Bee optimization based random double adaptive whale optimization model for task scheduling in cloud computing environment*, Computer Communications, 187 (2022), pp. 35–44.

[24] M. Masdari, F. Salehi, M. Jalali, and M. Bidaki, *A survey of pso-based scheduling algorithms in cloud computing*, Journal of Network and Systems Management, 25 (2017), pp. 122–158.

[25] ———, *A survey of pso-based scheduling algorithms in cloud computing*, Journal of Network and Systems Management, 25 (2017), pp. 122–158.

[26] R. Medara and R. S. Singh, *Energy efficient and reliability aware workflow task scheduling in cloud environment*, Wireless Personal Communications, 119 (2021), pp. 1301–1320.

[27] S. K. Mishra and R. Manjula, *A meta-heuristic based multi objective optimization for load distribution in cloud data center under varying workloads*, Cluster Computing, 23 (2020), pp. 3079–3093.

[28] A. Mohammadzadeh, M. Masdari, and F. S. Gharehchopogh, *Energy and cost-aware workflow scheduling in cloud computing data centers using a multi-objective optimization algorithm*, Journal of Network and Systems Management, 29 (2021), pp. 1–34.

[29] N. Mohanapriya, G. Kousalya, P. Balakrishnan, and C. Pethuru Raj, *Energy efficient workflow scheduling with virtual machine consolidation for green cloud computing*, Journal of Intelligent & Fuzzy Systems, 34 (2018), pp. 1561–1572.

[30] A. A. Nasr, N. A. El-Bahnasawy, G. Attiya, and A. El-Sayed, *Cost-effective algorithm for workflow scheduling in cloud computing under deadline constraint*, Arabian Journal for Science and Engineering, 44 (2019), pp. 3765–3780.

[31] M. Nematpour, H. Izadkhah, and F. Mahan, *Enhanced genetic algorithm with some heuristic principles for task graph scheduling*, The Journal of Supercomputing, 79 (2023), pp. 1784–1813.

[32] T. Nguyen, K. Doan, G. Nguyen, and B. M. Nguyen, *Modeling multi-constrained fog-cloud environment for task scheduling problem*, in 2020 IEEE 19th international symposium on network computing and applications (NCA), IEEE, 2020, pp. 1–10.

[33] A. Ramathilagam and K. Vijayalakshmi, *Workflow scheduling in cloud environment using a novel metaheuristic optimization algorithm*, International Journal of Communication Systems, 34 (2021), p. e4746.

[34] M. H. Shirvani, *A hybrid meta-heuristic algorithm for scientific workflow scheduling in heterogeneous distributed computing systems*, Engineering Applications of Artificial Intelligence, 90 (2020), p. 103501.

[35] P. Singh, M. Dutta, and N. Aggarwal, *A review of task scheduling based on meta-heuristics approach in cloud computing*, Knowledge and Information Systems, 52 (2017), pp. 1–51.

[36] A. Taghinezhad-Niar, S. Pashazadeh, and J. Taheri, *Energy-efficient workflow scheduling with budget-deadline constraints for cloud*, Computing, (2022), pp. 1–25.

[37] X. Wei, *Task scheduling optimization strategy using improved ant colony optimization algorithm in cloud computing*, Journal of Ambient Intelligence and Humanized Computing, (2020), pp. 1–12.

[38] A. M. Yadav, K. N. Tripathi, and S. Sharma, *An enhanced multi-objective fireworks algorithm for task scheduling in fog computing environment*, Cluster Computing, (2022), pp. 1–16.

[39] X. Zhang, T. Wu, M. Chen, T. Wei, J. Zhou, S. Hu, and R. Buyya, *Energy-aware virtual machine allocation for cloud with resource reservation*, Journal of Systems and Software, 147 (2019), pp. 147–161.

Edited by: SCPE Editor-in-chief
Regular paper
Received: May 17, 2024
Accepted: Dec 13, 2024

AIMS AND SCOPE

The area of scalable computing has matured and reached a point where new issues and trends require a professional forum. SCPE will provide this avenue by publishing original refereed papers that address the present as well as the future of parallel and distributed computing. The journal will focus on algorithm development, implementation and execution on real-world parallel architectures, and application of parallel and distributed computing to the solution of real-life problems. Of particular interest are:

Expressiveness:

- high level languages,
- object oriented techniques,
- compiler technology for parallel computing,
- implementation techniques and their efficiency.

System engineering:

- programming environments,
- debugging tools,
- software libraries.

Performance:

- performance measurement: metrics, evaluation, visualization,
- performance improvement: resource allocation and scheduling, I/O, network throughput.

Applications:

- database,
- control systems,
- embedded systems,
- fault tolerance,
- industrial and business,
- real-time,
- scientific computing,
- visualization.

Future:

- limitations of current approaches,
- engineering trends and their consequences,
- novel parallel architectures.

Taking into account the extremely rapid pace of changes in the field SCPE is committed to fast turnaround of papers and a short publication time of accepted papers.

INSTRUCTIONS FOR CONTRIBUTORS

Proposals of Special Issues should be submitted to the editor-in-chief.

The language of the journal is English. SCPE publishes three categories of papers: overview papers, research papers and short communications. Electronic submissions are preferred. Overview papers and short communications should be submitted to the editor-in-chief. Research papers should be submitted to the editor whose research interests match the subject of the paper most closely. The list of editors' research interests can be found at the journal WWW site (`http://www.scpe.org`). Each paper appropriate to the journal will be refereed by a minimum of two referees.

There is no a priori limit on the length of overview papers. Research papers should be limited to approximately 20 pages, while short communications should not exceed 5 pages. A 50–100 word abstract should be included.

Upon acceptance the authors will be asked to transfer copyright of the article to the publisher. The authors will be required to prepare the text in $\LaTeX 2_\varepsilon$ using the journal document class file (based on the SIAM's `siamltex.clo` document class, available at the journal WWW site). Figures must be prepared in encapsulated PostScript and appropriately incorporated into the text. The bibliography should be formatted using the SIAM convention. Detailed instructions for the Authors are available on the SCPE WWW site at `http://www.scpe.org`.

Contributions are accepted for review on the understanding that the same work has not been published and that it is not being considered for publication elsewhere. Technical reports can be submitted. Substantially revised versions of papers published in not easily accessible conference proceedings can also be submitted. The editor-in-chief should be notified at the time of submission and the author is responsible for obtaining the necessary copyright releases for all copyrighted material.