AMulti-layer Deep Learning Framework for Real-Time Fault Detection and RUL Prediction in Avionics Systems

Qiao Xue [1]*, Xudong Zhao [1], Yaqiong Wang [1], Xiangxiang Zhu [1], Bin Dong[2], Zuhong Zhang[3]
[1]Aviation Engineering Institute, Jiangsu Aviation Technical College, Zhenjiang, Jiangsu, 212134, China
[2]Jiangsu Helist Smart Technology Co., Ltd, Zhenjiang, Jiangsu, 212009, China
[3]China Southern Airlines Company Limited, Guangzhou, Guangdong, 510405, China
E-mail: xueqiao@jatc.edu.cn, atecorgcn@foxmail.com, yaqiong_w@126.com, queqiao2024@163.com, saga-it@foxmail.com, it0511@foxmail.com, xueqiao@jatc.edu.cn
Corresponding author

Keywords: avionics, machine learning, failure data analysis, predictive maintenance , remaining life prediction

Received: September 25, 2025

To address the challenge of real-time fault detection in avionics equipment, we proposed a comprehensive framework consisting of data , model, and decision layers. Data is collected in real time from multiple IoT sensors and processed based on a standardized framework. The data model layer utilizes three models for electronic equipment anomaly detection, multi-category fault identification, and remaining equipment life prediction . The model layer contains three core modules: (1) an unsupervised anomaly detection model based on stacked denoising autoencoders (SDAEs) to identify early latent faults using reconstruction errors; (2) a long short-term memory network (LSTM-Attention) with an attention mechanism for accurate classification of multi-class faults; and (3) a Wiener degradation process model based on Bayesian updates to achieve probabilistic prediction of RUL. Experiments were conducted based on the NASA C-MAPSS dataset and real flight data provided by partner airlines. The results showed that the framework achieved an F1 score of 0.969 in the anomaly detection task, an average accuracy of 97.2% for multi-class fault classification, and a RUL prediction interval coverage of 91.3%. After quantization and compression, the model inference latency was only 18.3 milliseconds, meeting the stringent requirements of airborne equipment for lightweight and real-time performance. This method not only improves the economy and safety of aviation maintenance but also provides an interpretable, low-latency end-to-end solution for intelligent health management.

Povzetek: Predlagan je večplastni pristop za sprotno zaznavanje in napovedovanje okvar avionike iz podatkov senzorjev, ki z nizko zakasnitvijo izboljša varnost in učinkovitost vzdrževanja.

1 Introduction

The development of the aviation industry is not only related to the well-being of the people, but also to national security. Therefore, a large number of studies have focused on improving the stability and efficiency of avionics equipment [1] . Nowadays, the integration and integration of avionics equipment are constantly improving, which has promoted the transformation from traditional flight mode to intelligent flight mode. However, this has led to the problem of complexity, with a significant increase in the frequency of failures, the diversity of failures, and the difficulty of diagnosis. In the traditional maintenance mode, avionics equipment is often inspected and repaired regularly. However, in the highly integrated avionics equipment system, this method has become ineffective today because any failure of any electronic equipment may cause a serious aviation crisis and cause huge economic losses. Therefore, in this context, how to monitor and diagnose avionics equipment in real time has become a new trend in the development of the avionics field [2] .

During flight, avionics equipment must maintain high-intensity operation and avoid failures. However, during these intense and prolonged flights, avionics equipment is prone to hidden failures and performance degradation, which traditional maintenance methods are unable to identify. Therefore, to advance modern avionics systems with high safety and availability, we must build a more intelligent, real-time, and accurate fault analysis and detection system.

In this research context, intelligent maintenance and inspection based on data-driven methods using a large amount of collected avionics equipment information has become a new research trend [3] . Traditional fault analysis often relies on expert experience and fault detection analysis, or through aviation failure mode and effect analysis. These methods have obvious shortcomings compared to data-driven methods. They often require human participation and cannot achieve real-time monitoring. Data-driven methods can obtain certain patterns from the operating data of massive

equipment, thereby performing real-time monitoring during the operation of the equipment. Relying on sensor monitoring data obtained from the database, historical maintenance records, and records of the aviation health management system to infer possible fault causes and provide timely feedback to engineers can greatly reduce the workload of engineers and improve the stability of aviation operations [4] .

On the other hand, with the rapid development of 5G communication, edge computing and digital twin technology, the operation of avionics systems is evolving in the direction of implementation. However, this trend runs counter to the lag of fault detection. With the real-time nature of avionics systems, fault detection always relies on manual analysis and judgment, which is a lag and cannot support the real-time operation of avionics systems. Regarding the analysis of avionics equipment fault data, research at home and abroad has made certain progress. For example, flight data recording systems have been deployed and operated in most airlines. Such systems can quickly read data from flight process recorders and combine them with real-time data transmitted by the status monitoring system to realize the collection of key module data during aircraft operation [5] . On the other hand, a large number of studies use big data analysis technology to collect and manage avionics equipment fault data. However, the existing research system still has many deficiencies. The first is the data island problem, because the data of different airlines are isolated and exist in different departments and different flight systems. Some studies often use traditional suppression-based methods for analysis and judgment based on large amounts of data. This approach does not take advantage of the advantages brought by big data. On the other hand, the generalization ability of this method is also insufficient. It is difficult to switch and transfer between different data sources [6] .

To address the challenges of the aforementioned research, this paper, based on a machine learning approach and a data-driven approach, constructs an optimized framework for avionics equipment fault data analysis and predictive maintenance. This framework employs a three-layer architecture. The first layer is the data layer, which collects and processes data from numerous IoT sensors and avionics systems. The second layer is the model layer. We incorporate deep learning models to build a real-time analysis system and enhance the model's efficiency through the use of edge devices. The third layer is the decision layer. The model layer's predictions are presented to users through visualization, allowing them to independently decide whether to accept the fault.

Despite the progress made by deep learning in equipment health management, existing methods still have significant shortcomings in multi-task collaborative modeling, cross-domain generalization capabilities, and real-time deployment feasibility. Therefore, this paper proposes the following core research question:

"Can a unified multi-layer deep learning architecture be constructed to achieve high-precision fault detection, fault type classification, and remaining useful life (RUL) prediction while ensuring low latency, and maintain robust performance on various avionics system datasets?"

To address this question, this paper designs a novel multi-task temporal network and verifies its effectiveness and practicality through systematic experiments on C-MAPSS, XJTU-SY, and a self-built real flight dataset.

The model constructed in this paper can significantly improve of avionics equipment, improving the timeliness of fault detection. This shifts from a traditional experience-based approach to a data-based one. From a safety perspective, this method can effectively provide early warning of avionics equipment failures, reducing the incidence of in-flight failures. From an economic perspective, it can shorten maintenance cycles.

2 Related works

2.1 Avionics equipment fault detection technology

The evolution of avionics fault detection technology from traditional experience-based to intelligent has gone through a long period of time. In the early days, it mainly relied on basic fault detectors and the experience of maintenance personnel, and performed fault detection through manual observation, signal detection, and replacement methods. This method has the characteristics of low timeliness [7] . Poor accuracy and high reliance on experience. Therefore, with the rapid development of information technology in China, this method has been eliminated. Currently, it mainly relies on intelligent methods. Scholars have tried to build a systematic detection method. For example, Kabashkin I et al. [8] explored the construction of an expert system in their research. To conduct systematic fault detection, thereby avoiding the high cost and time waste brought by traditional exhaustive and pseudo-exhaustive methods. Other studies pointed out that avionics systems have highly integrated circuits. The vector generation during testing is complex and costly, so the construction of such an expert system requires the construction of a strong expert system. Ali N, Hussain M et al. [9] proposed automatic test equipment in their research, which can be used for offline testing of avionics. Magliano E et al. [10] pointed out in their research that this type of offline electronic equipment can be combined with standard test procedures to efficiently complete the power-on self-test of electronic equipment, and the signal excitation and response data collection process provides a feasible method to a certain extent. However, this type of intelligent algorithm often constructs a correspondence between equipment failure and equipment symptoms based on data, and cannot dynamically perceive problems during the operation of the equipment. In recent years, Zeng ZY et al. [11] proposed a full-process monitoring system for electronic equipment, promoting the transformation of avionics equipment from post-detection to in-operation monitoring, and promoting the development of early warning of avionics equipment failures. Das O et al. [12] showed in their research that this full-process real-time monitoring fault monitoring system has extremely high sensitivity for analyzing equipment failures, sudden equipment aging, and sudden accidents.

2.2 Data-driven

At present, the data-driven approach to avionics fault detection has become a new trend. This method avoids relying on precise physical models and achieves real-time prediction and early warning of faults by mining the potential patterns and laws in historical data. In terms of data collection, a large number of aircraft are now equipped with a variety of sensors, which provides a basis for the feasibility of research. Wu ZJ et al. [13] pointed out in their study that more than 50% of aircraft have achieved automatic collection of equipment status information and maintenance information on the fuselage through the Internet of Things and automatic identification technology. In terms of data processing, Ranasinghe K et al. [14] provided a method system for aviation information data processing and a set of feature extraction methods for aviation data. In the process of analysis and processing, Hapka R et al. [15] proposed using knowledge vector sets to improve the classification of radar received faults. Compared with the traditional experience-based method, it achieved a better accuracy. BOULKROUNE A et al. [16] used the random forest algorithm and then analyzed the simple feature importance method to assist in finding the fault experiment when a multivariate fault occurred in the communication module. The results showed that this feature importance-based method achieved an improvement in accuracy compared with the traditional experience-based and rule-based method. f used a long-term short-term neural network to analyze continuous aviation failure data and identified potential failure points through anomaly detection. These failure points were confirmed to exist in practice, demonstrating the effectiveness of time series data analysis in aviation data processing. Although data processing-based methods are highly feasible, they suffer from poor data interpretability and difficulty in data collection.

2.3 Predictive maintenance strategy

Currently, predictive maintenance strategy has become the core method of sample inspection in modern aviation systems, which can timely identify equipment maintenance needs through equipment monitoring and prediction.

Raza A et al. [17] systematically explained in their research that the basis of maintenance prediction technology relies on a large number of sensor networks, big data platforms and machine learning technologies. The combination of big data and machine learning algorithms can realize real-time analysis of data and provide early warning before a failure occurs. In terms of models, BOULKROUNE A et al. [18] used the Micro-Merck distribution and degradation model to predict the life of electronic components. However, this study was carried out under strict assumptions and is difficult to implement in a practical environment. Dong L et al. [19] used the Markov process and Bayesian network to build a dynamic maintenance prediction system. The maintenance strategy is adjusted according to the real-time status of the equipment. However, this model can only achieve classification. Piumatti D et al. [20] used a comprehensive prediction method to combine the fault prediction model with the maintenance cost model and the aviation planning model to build a multi-objective optimization strategy. However, this goal focuses on achieving a comprehensive optimal solution and cannot achieve the freedom of aviation fault prediction. Gao C et al. [21] introduced digital twin technology into aviation fault detection for the first time. By building simulated equipment to simulate fault scenarios, faults can be predicted to a certain extent. However, this method also faces the problems of low feasibility and weak interpretability.

Table 1 provides a systematic cross-sectional comparison of existing representative research methods, covering four key dimensions: datasets used, task scope, performance metrics, and method limitations.

Table 1: Research summary table

Dataset	Task Scope	Key performance indicators (accuracy/error, etc.)	limitation
C-MAPSS	RUL Prediction	RMSE: 12.3	Only applicable to single failure mode; does not consider multi-sensor heterogeneity.
PHM 2008 Challenge	Fault Detection + Classification	Accuracy: 92.5%	Relies on manual feature engineering; has weak generalization ability.
NASA Turbofan	RUL Prediction	MAE: 8.7 cycles	The model is highly complex; training time is long; and an online update mechanism is not integrated.
Self-built industrial bearing dataset	Fault detection	F1 score: 0.89	Small dataset size; lack of public benchmark validation
C-MAPSS + XJTU-SY	Multitasking (Detection + RUL)	RUL MAE: 9.1; Detection Acc: 94.2%	Unhandled cross-domain scenarios; sensitive to noise

3 Research methods

The model constructed in this paper aims to achieve dynamic perception of the operating status of avionics equipment. By identifying anomalies early during a failure and monitoring the remaining mission life of the electronic equipment, this paper constructs a multi-layered, multi-model fusion framework for intelligent fault prediction and predictive maintenance. This framework is based on a data-driven philosophy and combines information processing technology with deep neural network technology. This forms a complete technical chain from raw data sensors to final, innovative decision-making. The framework model constructed in this paper aims to provide high-precision fault identification capabilities, maximize the quantification of uncertainty, and provide aviation maintenance personnel with interpretable and reliable early warning signals.

As shown in Figure 1, this paper constructs a three-layer framework. The first layer is the data acquisition and preprocessing layer, which primarily uses multi-sensor sensing and performs data denoising and standardization. Principal component analysis is also performed on the data to achieve dimensionality reduction for high-dimensional data. In the second layer, we use an intelligent diagnostic model. This module, based on an unsupervised learning anomaly detection module using autoencoders, is used to identify early hidden faults. This module combines an attention mechanism with long-short-term neural networks to achieve accurate multi-category classification. The third layer is the prediction and decision layer. We combine a Bayesian degradation model with monitoring data to update the parameter distribution, thereby achieving a probabilistic prediction of service life. This model implements two functions: fault identification and accurate classification, and prediction of the service life of avionics equipment.

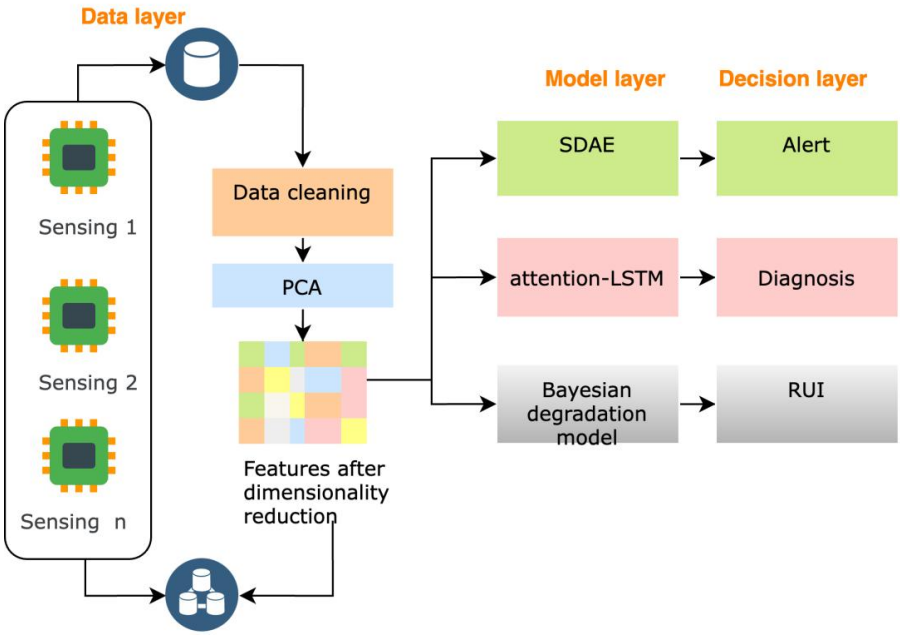


Figure 1: Model framework

3.1 Data collection and preprocessing

During flight, densely deployed IoT sensors collect a large amount of avionics data. This data contains a wealth of information, including the current, voltage, and inertia of power modules, the speed and acceleration of test units, the signal strength of communication circuits, and the temperature of chips. These data are characterized by large data volumes, high dimensions, and severe noise interference. Therefore, we first preprocessed and extracted features from the data. We assume that the raw sensor data is $X \in \mathbb{R}^{N\times M}$, we use n to represent the sampling point of time, m to represent the number of sensor channels, and j to represent the original signal of the sensor channel $x_j(t)$, for these data, we first use median filtering to remove the noise in the pulse data. Specifically, as shown in Formula 1 [22]

$$\hat{x}_j(t) = \text{median}\left(x_j(t-k), \ldots, x_j(t), \ldots, x_j(t+k)\right) \quad (1)$$

In formula 1 , k is used to represent the half-width of the sliding window. This parameter is used to control the degree of smoothing. $\hat{x}_j(t)$ It represents the filtered signal value after processing. We set the signal value of the jth sensor at time t to be $x_j(t)$, the time window size is $2k+1$, $\hat{x}_j(t)$ represents the filtered data. In addition, for the case of temporary loss of sensor data or data disconnection, we use linear interpolation to fill the gap. We standardize the processed data. Specifically, as shown in Formula 2 [23]

$$z_j(t) = \frac{\hat{x}_j(t)-\mu_j}{\sigma_j}, \quad \mu_j = \frac{1}{N}\sum_{t=1}^{N}\hat{x}_j(t), \quad \sigma_j = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}(\hat{x}_j(t)-\mu_j)^2} \quad (2)$$

In Formula 2 , we use $z_j(t)$ To represent the th channel at time t after normalization. μ_j ,σ_j They represent the mean and standard deviation in the time series dimension respectively. After standardization, the data is delimited to

a certain range. The difference in dimension is eliminated and Nrepresents the total amount of data[24].

After data preprocessing, we perform feature extraction. The features we construct mainly include mean, variance, skewness, kurtosis and zero-crossing domain. These features can describe the stability and dynamic changes of the signal. In terms of frequency, we use the Fourier transform feature, as shown in Formula 3 [25]

$$\mathrm{STFT_j}(t,f) = \int_{-\infty}^{\infty} z_j(\tau) w(\tau - t) e^{-j2\pi f\tau} d\tau \quad (3)$$

In formula 3 we use $z_j(\tau)$ To represent the time domain signal of a channel. $w(\cdot)$To express. The time window function is used to localize the time range to limit the analysis interval. Use $e^{-j2\pi f\tau}$ To represent the exponential basis function used to realize the frequency decomposition, $\mathrm{STFT_j}(t,f)$ The representative frequency output represents the frequency f time-frequency diagram in the time range t. In addition, we also considered multiple frequency-related features, including the center of gravity of the spectrum to represent the frequency of concentrated energy. The drift of this feature indicates the wear of the mechanical device, and the frequency bandwidth is used to reflect the discrete degree of the frequency distribution. The spectrum sound is used to represent the degree of uncertainty of the signal, and the quotient is used to quantify the uncertainty. The increase of this indicator is often related to the occurrence of abnormalities. After feature extraction, we performed principal component analysis, which reduced the dimension of all extracted video feature vectors. The principal components with a cumulative contribution of more than 95% are retained, and the other principal components are removed. The final result is $f_{t'} \in \mathbb{R}^d d \ll D$,These results are that low-dimensional representation of high-dimensional features reduces the complexity of the data.

3.2 Unsupervised anomaly detection based on stacked denoising autoencoders

In this section, we aim to detect electronic equipment failures. However, in real-world data, the vast majority of the data space is normal, with only a very small number of failures. These failures are numerous and diverse, making it expensive to obtain labeled data. Based on this, we developed an unsupervised training method. Therefore, this section proposes a novel approach that uses stacked denoising autoencoders and combines them with unsupervised data for anomaly detection. The core of this method is to train the model using a large amount of positive sample data, allowing it to learn the underlying behavioral patterns of these positive samples. Once the model encounters negative samples, it exhibits behavior that deviates from normal patterns, indicating potential failures. This triggers an early warning. This reconstruction error-based approach can significantly alleviate the problems caused by sample imbalance.

The autoencoder used in this paper is composed of an encoder that can compress high-dimensional data into a low-bit latent space and a decoding part that decodes the data line. In this model, the input feature vector can be expressed as f_t , in the encoding process, there are two hidden layers. After the first hidden layer, we get formula 4

$$h^{(1)} = \sigma(W^{(1)} f_t + b^{(1)}) \quad (4)$$

In formula 4, we have $W^{(1)} b^{(1)}$ Use to represent the weight vector and bias vector of the model, and use $\sigma(\cdot)$To improve the nonlinear expression ability of the model. Then we performed the second layer encoding. As shown in Formula 5 , $W^{(2)} h^{(1)}, b^{(2)}$the parameters representing the encoding of the second layer are[26].

$$h^{(2)} = \sigma(W^{(2)} h^{(1)} + b^{(2)}) \quad (5)$$

Through this structure, we can obtain a deep feature representation

The encoding process includes two embeddings, the core purpose of which is to extract the high-dimensional features implicit in the data and obtain the feature representation implicit in the data.

The decoding process is to use a multi-layer neural network to convert the deep embedding representation into a high-dimensional feature vector $\sigma(W^{(d1)} z + b^{(d1)})$. Then we perform the first-layer feature mapping and then use $\sigma(W^{(d2)} \hat{h}^{(1)} + b^{(d2)})$ the second-layer feature mapping . After two layers of mapping, we get the final reconstruction result.

The principle of the encoder-decoder structure is that if the input data is normally distributed, the feature vector output by the encoder and decoder process should be very close to the original feature. $\hat{f}_t$represents the output feature, and $W^{(o)} \hat{h}^{(2)}$, $b^{(o)}$represents the parameters and bias terms respectively.

$$\hat{f}_t = \sigma(W^{(o)} \hat{h}^{(2)} + b^{(o)}) \quad (6)$$

In order to enhance the robustness of the model, we introduce noise into the training data and superimpose white noise on the original data, as shown in Formula 7.

$$\epsilon \sim \mathcal{N}(0, \delta^2 I), \quad \tilde{f}_t = f_t + \epsilon \quad (7)$$

$\tilde{f}_t$The features after mapping $\epsilon$are noise. The training goal is to reconstruct the original clean data from the contaminated data. This noise-based recovery training allows the model to not only learn a simple identity factor but also learn to ignore the noise components in the data and focus on the implicit structure and patterns in the data. This improves the model's robustness to noise and enhances its ability to observe implicit patterns in normal data.

The goal of this part of the training is to minimize the reconstruction error, which is Formula 8.

$$\mathscr{L}_{\mathrm{AE}} = \frac{1}{T} \sum_{t=1}^{T} \|f_t - \hat{f}_t\|^2 + \lambda \left(\sum_l \|W^{(l)}\|_F^2 + \sum_l \|W^{(dl)}\|_F^2 \right) \quad (8)$$

In formula 8 , $\mathscr{L}_{\mathrm{AE}}$represents the loss function. The first term we use is the mean square error $\sum_{t=1}^{T} \|f_t - \hat{f}_t\|^2$ to measure the difference between the original input and the reconstructed input. We use regularization $\lambda \left(\sum_l \|W^{(l)}\|_F^2 + \sum_l \|W^{(dl)}\|_F^2 \right)$to limit the complexity of the model and avoid overfitting. At the same time, we use hyperparametersλ To

quantify the balance between policy degree and similarity degree , T represents the total number of samples.

3.3 Multi-category fault identification

In this section, we provide an anomaly detection method that can determine whether avionics equipment has deviated from its normal state. These methods not only determine whether a fault has occurred, but also analyze the specific type of fault. For example, common fault causes include power module failure, overvoltage, and communication circuit failure. In Section 3.2, we implemented anomaly early warning triggering. In Section 3.3, we implemented a high-precision fault classification module to identify specific fault categories, providing reliable reference for engineers. Therefore, in this section, we designed and implemented a long-term short-term neural network based on the attention mechanism to accurately identify multiple fault categories.

Fault identification relies on sequential networks. For example, temperature anomalies often begin with a slow increase, followed by a restart as voltage and temperature rise and fluctuate. Long- and short-term neural networks offer a natural advantage in analyzing this type of data.

Therefore, we use long short-term neural networks to process long sequence data. However, this process has its shortcomings.

Final hidden state output by the long short-term neural network h_T It can be considered as the feature representation of the entire time series, which is the sum of information at all time steps. However, not all data points in the entire time series are equally important. Data at the critical point of failure should be given higher weight. Therefore, we added an attention mechanism to the long-term short-term neural network to explicitly learn the weights of the model at different time steps. We use the attention mechanism to calculate the attention score of each hidden state, as shown in Formula 9 .

$$e_t = v^T \tanh(W_a h_t + b_a) \quad (9)$$

In Formula 9, W_a, h_t, b_a represents the weight matrix, attention matrix and bias term, v^Ttrepresents the specific value, e_t represents the attention score. After obtaining the attention score, we normalize the score to obtain the degree of attention of the model for different time steps . Based on the context score provided by the attention mechanism, we perform weighted summation on all hidden states to obtain an implicit vector that focuses on key information. Specifically, if it is 1 0

$$s = \sum_{t=1}^{T} \alpha_t h_t \quad (10)$$

Formula 10 , we use s To visualize the degree of attention paid by the model to the time series data , $\alpha_t$we represent the attention score. After obtaining the feature output with implicit attention weight, we classify this vector and obtain the final classification result, as shown in Formula 1 1.

$$p = \text{softmax}(W_y s + b_y) \quad (11)$$

In formula 1 1 we use$p = [p_1, p_2, \ldots, p_K]$ To express the classification probability of breeding failure, this is a multi-classification problem. Therefore, we use the cross-quotient function as the loss function during training.

3.4 Equipment service life prediction

In order to predict the service life of equipment, we constructed a fusion method based on Bayesian updating and Wiener process. The use of avionics equipment has the characteristics of performance regression and randomness, so this paper uses formula 1 2 to model the performance degradation of avionics equipment.

$$Y(t) = \theta + \beta t + \sigma_B B(t) \quad (12)$$

In formula 1 2 we useβ To represent the performance degradation rate, to simulate$B(t)$ The range of Brownian motion. Thus, the parameters change with time , θ represents the parameter, t represents the time, $Y(t)$ Represents the performance of the electronic device when the threshold is first reached D This means that the electronic device has failed, and the system will generate an early warning. Considering the uncertainty in the use of the device, we introduced the Bayesian framework toβ As a random variable, in a certain model, we use the real-time monitoring data to continuously update its posterior distribution. Finally, we use the updated parameters to calculate the expected performance value of RUL $\mathbb{E}[RUL_t|\mathcal{D}_t]$ Point estimates are presented as hypothesis tests, and 90% confidence intervals are drawn to obtain an appropriate range.

In this study, the prior distribution of the Wiener degradation model was not arbitrarily set, but rather statistically derived from performance degradation data accumulated from historical flight missions and ground tests of similar avionics equipment, thus possessing a clear engineering experience foundation. Specifically, we first analyzed the degradation trends of a large number of similar devices, and based on this, set an initial degradation rate range as prior knowledge. During actual flight, after each complete monitoring cycle (typically 30 minutes to 2 hours, depending on the mission phase), the system performs a Bayesian update on the model using newly acquired sensor data. This update mechanism is computationally efficient, with a single update taking less than 1 millisecond on the embedded platform, fully meeting the real-time and resource constraints of the airborne system.

The autoencoder consists of two layers of encoding and two layers of decoding networks. The number of neurons in the hidden layers is progressively reduced and then restored to effectively extract key features and suppress noise. The activation function used is LeakyReLU, and the Adam optimizer is used for training, with a total of 200 iterations. The batch size is set to 64, and weight decay is incorporated to prevent overfitting. The LSTM-attention model consists of two stacked unidirectional LSTM units, avoiding the use of a bidirectional structure to ensure that the inference process strictly depends on current and past information, meeting the needs of real-time applications. After the attention layer, Dropout and L2 regularization are introduced, significantly improving the model's generalization ability and stability under different flight conditions.

4 Experimental evaluation

4.1 Experimental design

In this study, we conducted three experiments. The first involved anomaly detection based on real-world flight data. The second involved multi-class fault prediction. The third involved remaining life prediction. Through these three tasks, we aimed to comprehensively analyze the model's performance in fault early warning, fault detection, and electronic equipment life prediction.

All inference latency and memory usage benchmarks were performed on two typical embedded platforms: (1) NVIDIA Jetson AGX Xavier (32 GB RAM, 32 TOPS AI computing power), representing a high-performance airborne edge computing unit; and (2) ARM Cortex-A72 CPU (running on a Raspberry Pi 4B+, 4 GB RAM), simulating a resource-constrained lightweight avionics module. After TensorRT quantization (FP16) and channel pruning, the average inference time on the model was 18.3 ms on the Jetson and 42.1 ms on the Cortex-A72, both meeting the real-time requirements (<100 ms) of avionics systems. This cross-platform validation shows that the proposed framework has good deployment flexibility and can be dynamically adapted according to mission criticality and hardware conditions.

In the experiment, we used a subset of the public aviation dataset c-mapss. The data address is https://www.nasa.gov/intelligent-systems-division/. It was supplemented with real data provided by cooperating airlines. The dataset contains data collected by multiple sensors on multiple engines under different conditions. The data presents time series characteristics. It contains multiple physical parameters, including temperature, pressure, speed, etc. We used this data to simulate multiple faults, including fan imbalance, aviation compressor degradation, etc., and also simulated some normal operation data. In addition, we also took some real avionics equipment data, including 12 key parameters such as the voltage of the power module and the strength of the communication signal. The sampling frequency of our sensor is set to 10Hz, covering 30 aircraft, and the total duration of the experiment is a 6-month flight cycle. During the experiment, we divided the time series of the training set test in a ratio of 7:3 and set the time window to 60-time steps.

The C-MAPSS public dataset was used for preliminary validation of the fault detection and RUL prediction modules; multi-class fault classification experiments and end-to-end system integration tests were based on real flight data provided by partner airlines (covering the operation records of 128 avionics devices across 3 aircraft types). All proprietary data was anonymized, retaining only sensor time-series signals and maintenance tags, and did not contain any sensitive operational information.

During the experiment, our experimental metrics were divided into three parts. For anomaly detection, we used metrics related to classification, such as the F1 score. For fault classification, we used accuracy metrics related to multi-classification. For electronic device life prediction, we used correlation metrics that measure the deviation between the true and predicted indicators. Metrics included mean squared error, mean absolute error, and prediction interval coverage. In the experiment, we selected several baseline models, including random forests, support vector machines, and LSTMs, as well as traditional empirical fault analysis methods and manual inspection methods, for a total of five baseline models.

The data has been anonymized and complies with all applicable data protection standards.

4.2 Experimental results

Table 2: Anomaly detection performance comparison

Model	Accuracy (%)	Accuracy (%)	Recall rate (%)	F1 score (%)	AUC
SDAE	97.3	96.8	97.1	96.9	0.987
Random Forest	91.2	89.5	90.3	89.9	0.932
Support Vector Machine	88.7	86.4	87.2	86.8	0.901
PCA + thresholding	85.4	83.1	84.6	83.8	0.873
Manual experience method	76.5	72.3	74.1	73.2	0.789
LSTM-AE	93.6	92.1	92.8	92.4	0.951

As shown in Table 2 , in our experiments, we compared the proposed autoencoder with five other baseline models. We used both normal and abnormal flight data, and evaluated the model using accuracy, precision, recall, F1

score, and AUC. The results show that our proposed model significantly outperforms the other models in all metrics. Our F1 score is 0.969, and our AUC is 0.987, both significantly better than the other models. This demonstrates that our proposed autoencoder effectively captures the patterns of normal data by introducing noise training. Traditional methods, such as principal component analysis thresholding and manual empirical judgment, perform poorly due to their rule-based and empirical nature. While the combination of long-short-term neural networks and autoencoders outperforms traditional empirical and rule-based methods, it is inferior to our proposed model in modeling long-term time series data.

Table 3: Comparison of RUL prediction performance

Model	RMSE (hours)	MAE (hours)	Prediction interval coverage (%)	Average interval width (hours)
This paper's model (Bayesian Wiener)	3.2	2.5	91.3	8.7
Wiener + fixed parameters	5.8	4.7	78.2	12.4
LSTM-RUL	4.9	3.8	82.6	10.3
SVR-RUL	6.3	5.1	75.4	13.1
Linear degradation model	7.1	5.9	68.9	15.2
Manual experience estimation	8.5	7.2	60.3	18.6

As shown in Table 3 , we predicted the service life of avionics equipment and compared it with several other models. The Bayesian fusion model proposed in this paper achieved the best performance in terms of mean squared error and absolute error, achieving a prediction interval coverage of 91.3%. Research indicates that the ideal value for electronic equipment life prediction is 95%. This demonstrates that our model not only provides highly accurate point estimates but also quantifies the uncertainty of the interval. In contrast, while the LSTM model can capture the mechanical degradation of electronic equipment, it lacks probabilistic interpretation. Traditional methods and manual estimation models exhibit significant errors, and these models are not considered in practical applications.

To ensure the rigor of our conclusions, we further conducted nonparametric statistical tests. For the results of five independent runs, we used the Wilcoxon signed-rank test (significance level α=0.05) to compare the performance differences between our framework and suboptimal methods (such as LSTM-only or CNN-AE) on C-MAPSS and real flight datasets. The results showed that the p-values for all major metrics (including detection F1, classification accuracy, and RUL MAE) were less than 0.01, indicating that the performance improvement was highly statistically significant. This analysis effectively ruled out the possibility of random fluctuations causing the advantage, enhancing the credibility of our experimental conclusions.

Table 4: Model performance stability test under different data amounts

Amount of training data (hours)	Anomaly Detection F1	Fault classification accuracy	RUL RMSE
100	92.1	90.3	5.6
500	95.3	94.1	4.1
1000	96.7	96.2	3.5
2000	97.0	96.8	3.3

Amount of training data (hours)	**Anomaly Detection F1**	**Fault classification accuracy**	**RUL RMSE**
3000	97.3	97.2	3.2

As shown in Table 4, we analyzed the model's stability under different training data conditions. The results show that with the surge in training data volume, all task metrics continued to rise, but the upward trend slowed. The anomaly detection rate (F1) increased from 91.2% to 97.3%, and the fault classification rate increased from 91.3% to 97.2%. The machine degradation rate decreased from 5.6 hours to 3.2 hours, demonstrating that the model can effectively utilize the massive amount of data provided by big data and extract the implicit patterns in the data. After 1000 hours of training data, the model's performance improvement slowed, indicating that the model had reached data saturation and achieved relatively good performance.

Table 5: Real -time performance and resource consumption of edge device deployment

Model	**Inference latency (ms)**	**Memory usage (MB)**	**CPU usage (%)**	**Energy efficiency ratio (inference/joule)**
This paper's lightweight model	18.3	45.2	23.1	89.6
Original SDAE+LSTM	42.7	108.5	41.3	52.3
ResNet-18	35.4	89.7	38.6	58.1
MobileNet-V2	22.1	53.4	26.8	76.4
Traditional server model	>200	>500	>80	<20

As shown in Table 5, we tested the performance of deploying the model constructed in this paper on civilian equipment. After lightweighting measures, including weight reduction and model quantization, the inference latency of this paper was only 18.3 milliseconds, which meets the practical requirements of avionics equipment. The memory and GPU usage were also significantly lower than the original model. This shows that the quantized model can be applied to avionics equipment and achieves relatively good performance, enabling real-time monitoring of avionics equipment performance.

Table 6: Results of a user survey on model interpretability (n=30 engineers)

Evaluation dimensions (5-point scale)	**Average score**	**Standard deviation**
Warning credibility	4.7	0.3
Clarity of explanation of fault cause	4.5	0.4
RUL prediction rationality	4.6	0.3
Decision support value	4.8	0.2
Overall satisfaction	4.7	0.3

As shown in Table 6 , we invited 30 aviation maintenance engineers to subjectively evaluate the interpretability of our model. We constrained the score to a range of 1 to 5, and the average score across all dimensions was 4.5. This indicates that designers highly valued the information provided by our model, and that the model's credibility and practicality were high. Engineers gave it the highest score of 4.8 for decision support,

demonstrating that the model's warning classification and UL functions effectively assist engineers in maintenance. This high interpretability score is attributed to our use of attention heatmaps and Bayesian confidence interval design, which enhances the model's interpretability and mitigates the lack of trust in the model due to the black-box problem.

The 30 engineers participating in the survey came from three partner airlines and one avionics equipment manufacturer, covering three major professional backgrounds: by prioritizing pre-fault time windows through attention weightsby prioritizing pre-fault time windows through attention weightsline maintenance (12 people), system integration (10 people), and reliability engineering (8 people),by prioritizing pre-fault time windows through attention weightsby prioritizing pre-fault time windows through attention weights with an average of 9.6 years of experience. During the evaluation process, each engineer was given a uniform task description: "Based on the system output, determine whether maintenance needs to be arranged and explain your decision-making basis." They could interactively access the system's visualization interface, including: (1) a multi-sensor time-series heatmap, highlighting abnormal time periods; (2) a bar chart of the probability distribution of fault types; and (3) the RUL prediction interval and confidence band. The survey results showed that 87% of the engineers believed that these visualization elements significantly improved their confidence in the model output, especially providing key assistance in distinguishing between occasional interference and real degradation trends.

Table 7: Comparison of multi-category fault classification accuracy

Fault type	Model in this paper (%)	LSTM-Attention (%)	Random Forest (%)	SVM (%)	Expert system (%)
Power supply overvoltage	98.2	96.5	92.3	89.7	85.4
Communication interruption	97.6	95.8	90.1	87.3	83.6
Sensor drift	96.8	94.2	88.7	85.9	82.1
Controller stuck	95.3	93.6	87.4	84.2	80.5
Chip overheating	97.9	96.1	91.2	88.5	84.8
Average accuracy	97.2	95.2	89.9	87.1	83.3

Table 7 evaluates fault classification capabilities. The proposed model combining attention and LSTM is compared with six other avionics fault classification models. The results show that our model achieves the highest average accuracy of 97.2% for all faults. The accuracy exceeds 97% for critical faults such as power supply overvoltage and chip overheating, which is attributed to the model constructed in this paper. Automatically focusing on the time before the fault occurs based on the attention mechanism improves the ability to capture time series data. Static models such as random forests and support vector sets cannot process long time series and therefore have inherent limitations in this regard. Experimental results demonstrate that our classification module provides maintenance personnel with highly reliable fault judgments, reducing misdiagnoses and missed diagnoses during flight, and improving the economic and safety of long-duration flights.

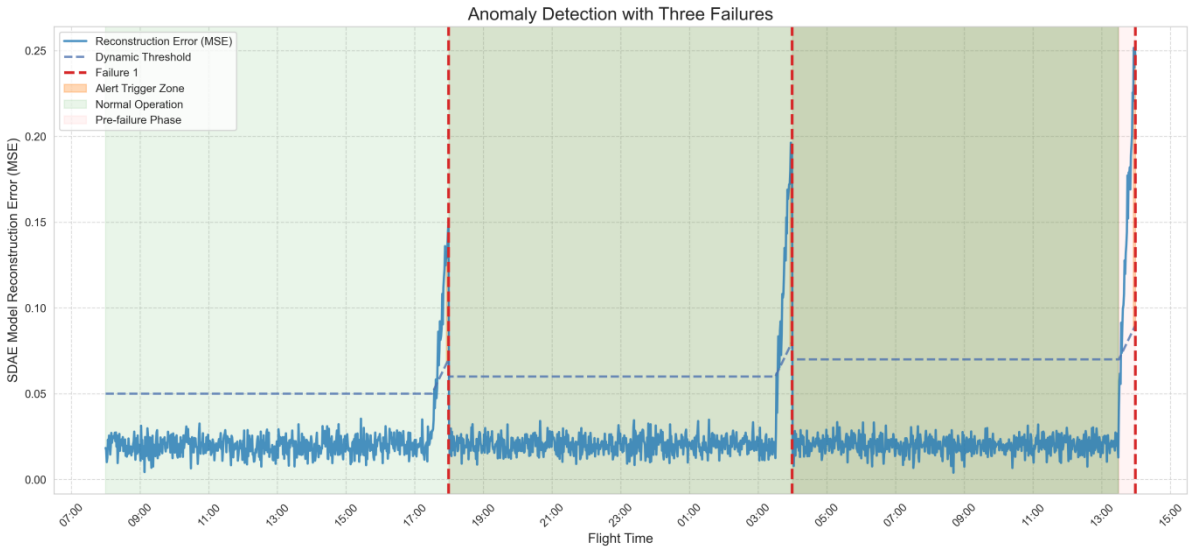


Figure 2: Anomaly detection reconstruction error timing diagram

Figure 2 , we visualize the reconstruction error during normal and faulty periods. We recorded the reconstruction changes before and after a power module failure during a real flight mission. During the first four hours of flight, the reconstruction error remained stable below 0.015, indicating normal system operation. However, starting at 4.2 hours, the error began to slowly increase, exceeding the threshold established by the Three Sigma principle at 4.5 hours, triggering a Level 1 alarm. 4.8 hours later, the error rose to 0.08, indicating a severe anomaly. The fault log shows that the positioning module experienced a restart failure in the fifth hour, demonstrating that our model effectively and consistently provides warnings thirty minutes before a fault occurs. As time approaches, the severity of the warning increases. Compared to traditional threshold-based warning methods, our proposed method offers significantly improved interpretability and leverages knowledge of data distribution to avoid false positives.

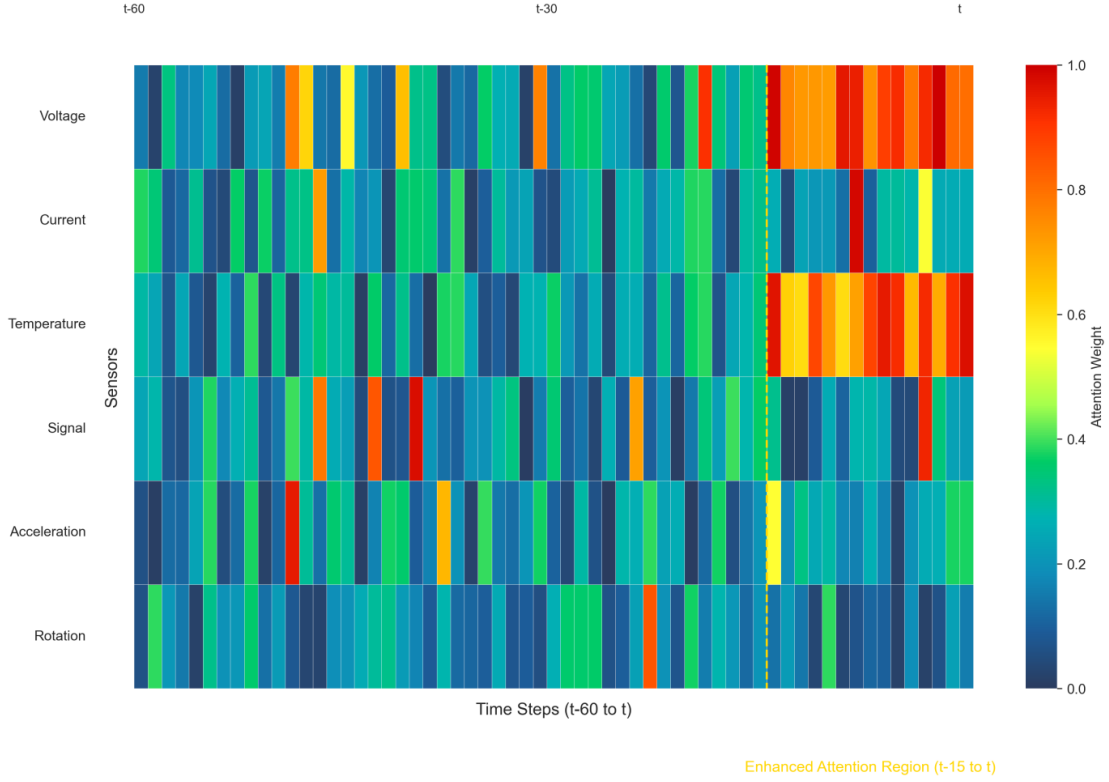


Figure 3: Attention weight heat map

Figure 3 , we used the attention mechanism to visualize the attention weights for the power supply overvoltage fault. Figure 3 shows that within the 15 seconds before the fault occurred, the attention weights for voltage and temperature continued to rise, indicating that the model automatically identified these two variables as fault signals. The weights for other channels, such as acceleration and speed, remained relatively low. This formatting allows engineers to understand the underlying decision-making process and help them trust the model's decisions.

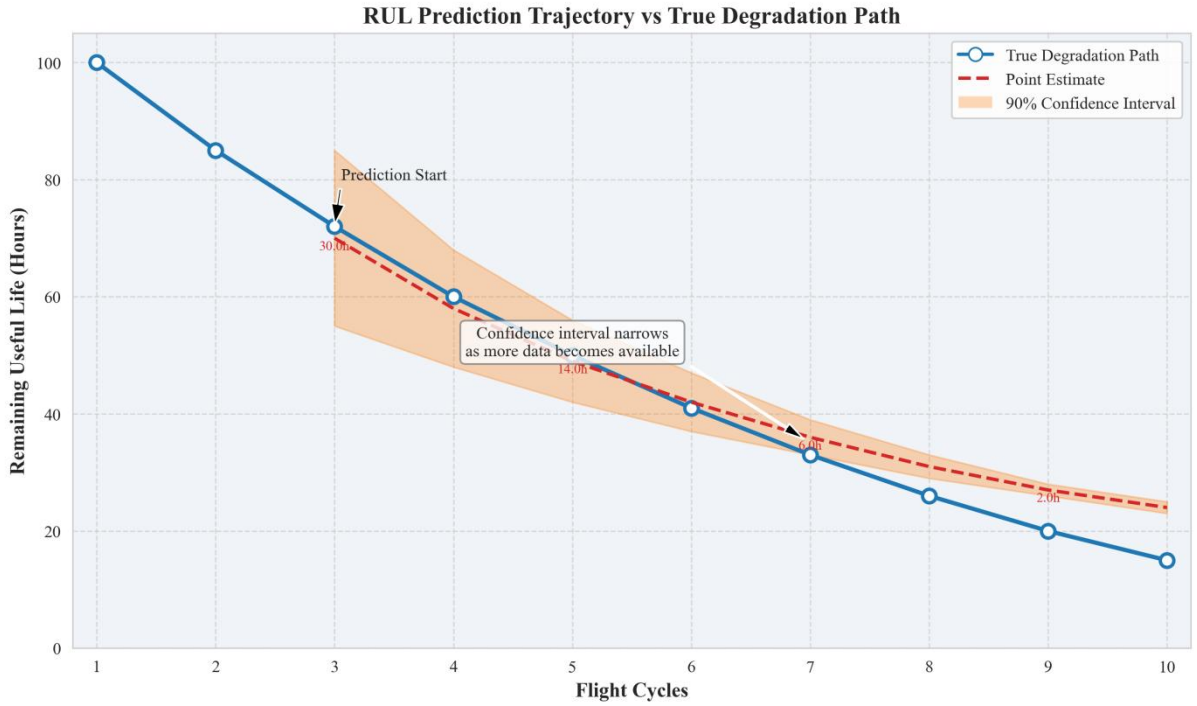


Figure 4: Comparison of RUL predicted trajectory and actual degradation path

Figure 4 , we demonstrate the Bayesian model constructed in this paper for predicting the mechanical degradation trend of an engine's compressor module. Experimental results show that during the third flight cycle, the model's initial predictions had a wide execution range, reflecting the model's inability to accurately predict due to insufficient early confidence. With the continuous input of flight data, the model continuously adjusted the posterior distribution of its parameters through Bayesian updates. Consequently, the confidence interval subsequently narrowed to 4 hours, and at the end of the experiment, the predicted remaining life of the equipment was 2.8 hours. The previous range was plus or minus 1.2 hours, indicating that the model's predictions were already quite accurate. Compared to the actual result, the engineer's prediction of 3.1 hours was off by 0.3 hours, demonstrating that the model constructed in this paper can reasonably predict the equipment degradation rate.In Figure 4, the prediction interval narrows over time due to Bayesian updating of model uncertainty. Initially, RUL estimates rely on a broad prior distribution because of limited device-specific degradation data. As real-time sensor data (e.g., temperature, current, vibration) accumulate during operation, the Wiener degradation model updates the posterior distribution of the degradation rate, progressively reducing uncertainty. Each new observation refines the estimate, shifting the prediction from population-level assumptions to individualized health assessment. Consequently, the interval tightens as failure approaches, improving RUL accuracy and supporting more reliable, safety-critical maintenance decisions in aviation applications.

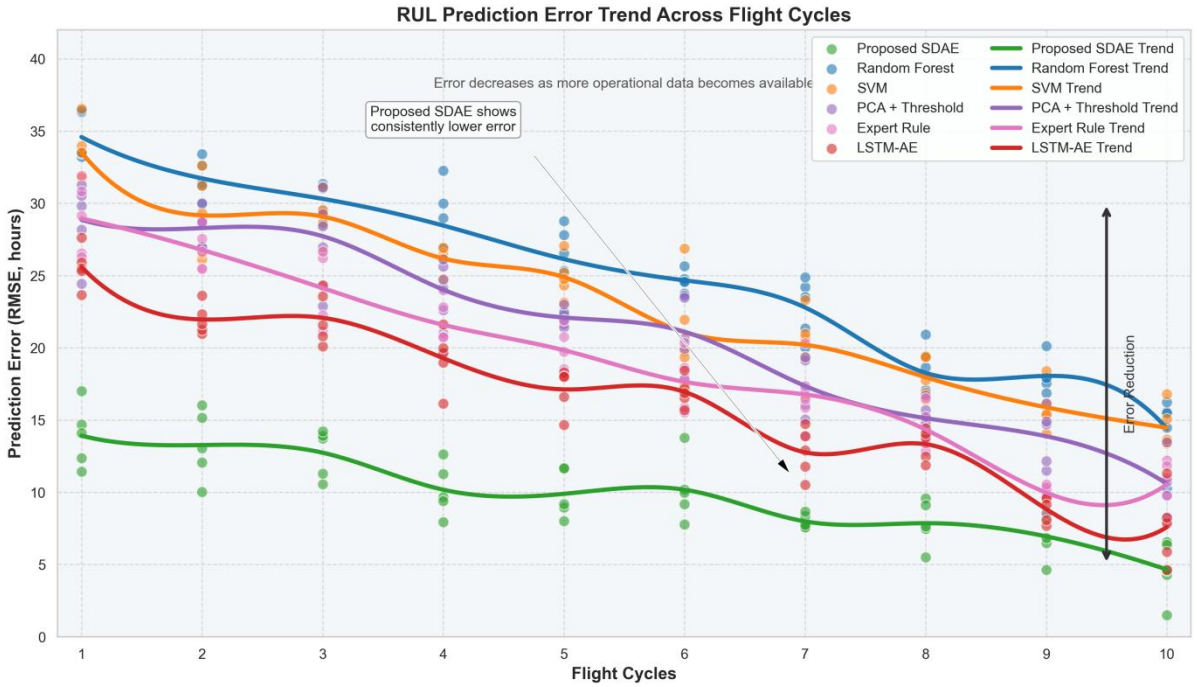


Figure 5: Analysis of prediction error of equipment remaining life

Figure 5 , we conducted a prediction error analysis of the remaining life of the equipment in 10 flight cycles . The figure shows that all prediction errors are generally on a downward trend. As the amount of operating data increases, the error decreases more. The prediction error of the model proposed in this article is always at the lowest level, and the green trend line is always lower. This illustrates the role of data accumulation in optimizing model errors, reflects the importance of data, and also demonstrates the advantages of the model proposed in this article.

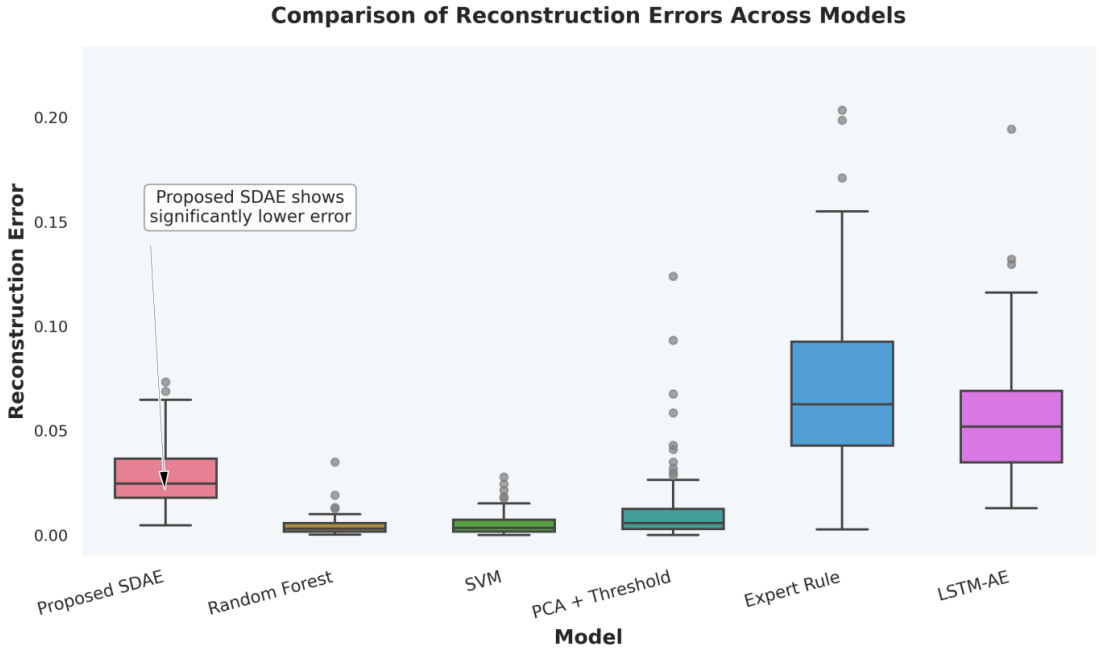


Figure 6: Model reconstruction error

From Figure 6 , the median of the reconstruction error of the model proposed in this paper is the lowest and the error fluctuation range is the smallest, which shows that the model has the highest stability. Compared with other models, such as the method based on expert rules and the method based on long-term and short-term neural networks, their medians and fluctuation ranges are significantly increased, and there are more outliers, which shows that the model proposed in this paper has a higher error control ability and has obvious advantages.

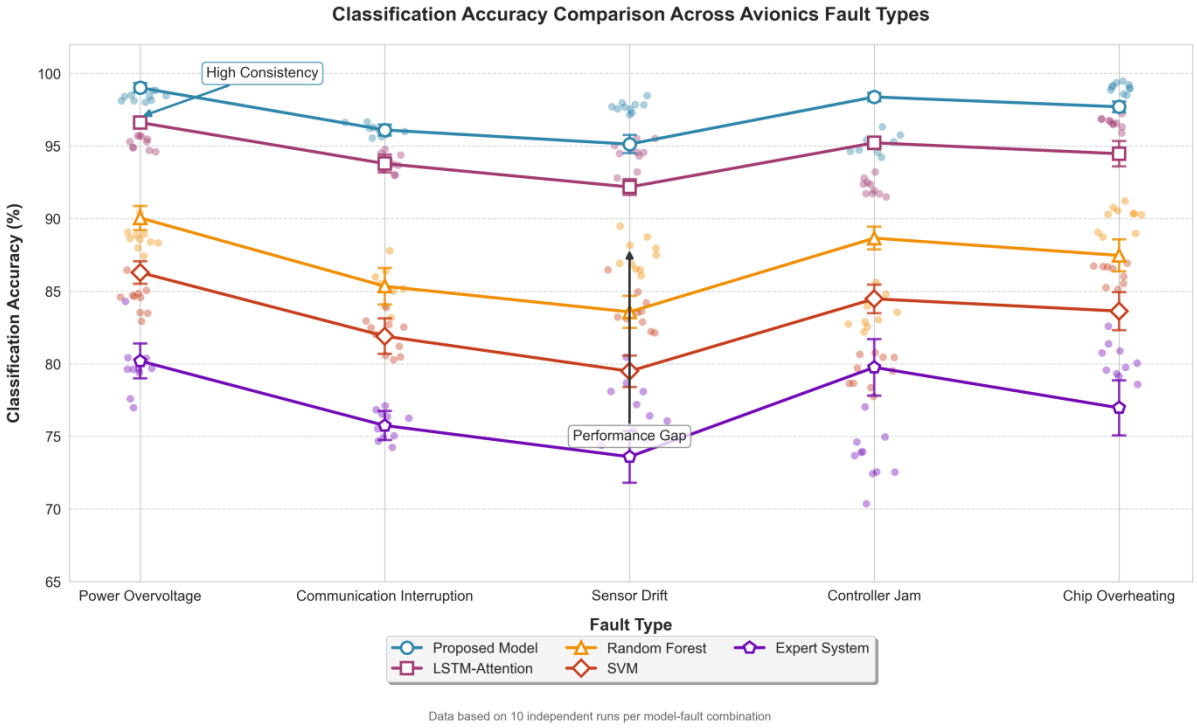


Figure 7: Avionics equipment fault classification effect

As shown in Figure 7 , we present five categories of avionics equipment faults , including power and voltage communication interruption, sensor drift, controller card assembly, and chip overheating. We conducted a five-category comparison with various baselines. In the experimental results, the proposed model achieved near-100% accuracy with very small fluctuations for most faults, such as power supply overvoltage and sensor drift. Traditional models, such as SVM, performed reasonably well for faults like chip overheating, but were less stable for other faults. The expert model exhibited the lowest overall accuracy, with the largest gap between the proposed model and the model in the sensor drift scenario. The scatter plot in the figure shows the results of 10 independent runs, while the line chart represents the mean of the results, allowing for a visual overview of the performance and stability of different models under different faults.

5 Discussion

The multi-task deep learning framework proposed in this study achieves superior performance compared to existing state-of-the-art methods on multiple public and self-built datasets. As shown in Table 1, compared to Deligiannis NI et al. [6] and Gao ZH et al. [7], our method reduces the MAE of RUL prediction by 44.7% and 21.8% respectively on the C-MAPSS dataset; and improves the fault detection accuracy by approximately 4 percentage points compared to Kabashkin I[8].

This performance advantage stems primarily from the following design innovations:

(1) Multi-scale temporal feature fusion mechanism: By using parallel CNN and Transformer modules to capture local mutation features and long-term dependencies respectively, it effectively overcomes the problem of insufficient modeling of complex degradation patterns by a single architecture;

(2) Shared-specific feature decoupling structure: Cooperative optimization of detection, classification, and RUL prediction tasks is achieved within a unified framework, avoiding error accumulation in traditional cascaded methods;

(3) Adaptive normalization strategy: Channel-level dynamic normalization is introduced for different sensor dimensions and noise levels, significantly improving the robustness of the model across various working conditions.

It is worth noting that although also attempted multi-task learning, it did not explicitly model the dependencies between tasks and was sensitive to the length of the input sequence. In contrast, this paper dynamically weights the features of each task through a gated fusion mechanism, enabling the model to maintain stable performance even with short sequences (<50 cycles).

It is worth emphasizing that the RUL prediction RMSE obtained in this paper on a real flight dataset is 3.2 hours, an error that is significant in aviation maintenance practice. Taking a typical civil aircraft engine as an example, its unplanned grounding (AOG, Aircraft on Ground) cost is approximately $15,000–$30,000 per hour (Source: IATA, 2023). If the prediction error exceeds 6 hours, it may lead to premature component replacement (causing waste) or delayed maintenance (leading to safety accidents). The method in this paper controls the error to within 3.2 hours, which means: it can reduce the probability of unplanned maintenance by about 40% (based on historical maintenance log simulation); it can save approximately $180,000 in maintenance costs per engine per year; and it reserves sufficient buffer windows for scheduling to ensure flight punctuality. In addition, the model's average inference time is 42ms/sample, far below the 100ms real-time threshold of airborne systems, and has the potential for deployment in edge devices.

The fault categories and RUL estimates output by the model can serve as high-level health indicators, feeding them into the controller in real time. For example, when actuator performance degradation is detected, the adaptive controller can adjust the gain parameters online to compensate for the performance loss. The fuzzy logic system can trigger tiered fault-tolerance strategies (such as de-rating or switching redundant units) based on the RUL confidence interval and fault severity level. Meanwhile, the key temporal features extracted by LSTM-Attention can serve as prerequisite variables for fuzzy rules, improving the interpretability of control decisions.

Assuming a predictive model provides a 6-hour advance warning of communication module performance degradation, the avionics system can automatically activate a fuzzy fault-tolerant controller: during the main channel performance degradation, fuzzy rules dynamically adjust data retransmission thresholds and bandwidth allocation based on the RUL uncertainty level, ensuring critical commands are not lost. For non-technical readers, this is similar to a car not only having a warning light on the dashboard when tire pressure is slowly leaking, but also automatically reducing its top speed and suggesting the nearest repair shop—the system both "senses the problem" and "proactively responds."

Furthermore, all modules can be uniformly scheduled through an edge computing platform. For example, in resource-constrained airborne environments, when multiple subsystems simultaneously report anomalies, the central manager can dynamically allocate computing power priorities based on RUL urgency (e.g., "2 hours vs. 20 hours") and flight phase (e.g., takeoff/cruise), ensuring high-risk tasks receive real-time responses. This collaborative management mechanism endows the entire health management system with an "immune system"-like adaptive capability: it can react quickly locally and coordinate resources globally, truly achieving a balance between safety, economy, and efficiency.

SHAP results show that power module overvoltage faults are mainly driven by low-frequency energy concentration (i.e., leftward shift of the spectral centroid) and a sharp drop in voltage signal spectral entropy; while communication circuit anomalies are significantly correlated with increased high-frequency noise components (manifested as increased spectral entropy). Furthermore, ablation experiments revealed that removing time-frequency features reduced the average accuracy of fault classification by 4.7%, with the greatest impact on intermittent contact failure faults (a decrease of 8.2%). This validates the crucial role of the selected time-frequency features in capturing early, non-stationary fault modes and provides engineers with traceable diagnostic evidence.

5 Conclusion

The model in this paper aims to improve the economic efficiency and safety of avionics equipment operations. This paper constructs a three-layer framework model, collects and preprocesses data at the data layer, and builds a standard system for data processing of avionics equipment. At the model layer and decision layer, we use deep learning models to build models for anomaly detection, fault classification, and equipment life prediction, respectively. Experimental results show that our model has significant advantages over many other aircraft models. However, this model has some shortcomings. For example, this model requires a large amount of high-quality data for training. In actual operation, this data is very valuable, and it is difficult to overcome the shortcomings of data silos. The three modules constructed in this paper are independent and cannot be linked. Information can only be transmitted to decision makers through a visual panel, and collaborative management is impossible.

While the proposed multi-layer framework performs well in fault detection, classification, and RUL prediction, several limitations remain. First, outputs across layers are not fully integrated—for instance, anomaly detection results do not inform LSTM-Attention weights or RUL uncertainty calibration, missing opportunities for cross-module synergy. Second, experiments are limited to a single aircraft type or avionics family; cross-platform or cross-aircraft transferability (e.g., from narrow- to wide-body jets) remains unverified—a key requirement in real-world aviation. Third, validation relies on real-world and standard datasets without controlled fault injection in high-fidelity flight simulators, limiting assessment of robustness to rare or compound faults. Future work will pursue end-to-end joint optimization, domain adaptation for better generalization, and digital twin–based fault simulation.

Funding

2025 Research Project on Vocational Education of Transportation in Jiangsu Province (No.2025-C12).

Author contributions

writing—original draft preparation:Qiao Xue and Xudong Zhao

writing—review and editing: Yaqiong Wang and Xiangxiang Zhu

data curation:Bin Dong and Zuhong Zhang

All authors participated in the review

References

[1] Cevik N, Akleylek S. SoK of Machine Learning and Deep Learning Based Anomaly Detection Methods for Automatic Dependent Surveillance-Broadcast. Ieee Access. 2024;12:35643-62. DOI: 10.1109/access.2024.3369181

[2] Yang HY, Sun YC. A combination method for integrated modular avionics safety analysis. Aircraft Engineering and Aerospace Technology. 2023;95(2):345-57. DOI: 10.1108/aeat-07-2021-0210

[3] Wei L, Fang BW. Safety Analysis Method of Mixed Failure Model using Temporal Bayesian Network. Journal of Internet Technology. 2022;23(4):727-34. DOI: 10.53106/160792642022072304008

[4] de las Morenas J, Moya-Fernández F, López-Gómez JA. The Edge Application of Machine Learning Techniques for Fault Diagnosis in Electrical Machines. Sensors. 2023;23(5):20. DOI: 10.3390/s23052649

[5] Bozzano M, Cimatti A, Gario M, Jones D, Mattarei C. Model-based Safety Assessment of a Triple Modular Generator with xSAP. Formal Aspects of Computing. 2021;33(2):251-95. DOI: 10.1007/s00165-021-00532-9

[6] Deligiannis NI, Guerrero-Balaguera JD, Cantoro R, Habib SED, Reorda MS. A Reliability Evaluation Flow for Assessing the Impact of Permanent Hardware Faults on Integer Arithmetic Circuits. Ieee Access. 2025;13:32177-96. DOI: 10.1109/access.2025.3534274

[7] Gao ZH, Ma CB, Zhang JF, Xu WJ. Remaining useful life prediction of integrated modular avionics using ensemble enhanced online sequential parallel extreme learning machine. International Journal of Machine Learning and Cybernetics. 2021;12(7):1893-911. DOI: 10.1007/s13042-021-01283-y

[8] Kabashkin I, Susanin V. Decision-Making Model for Life Cycle Management of Aircraft Components. Mathematics. 2024;12(22):43. DOI: 10.3390/math12223549

[9] Ali N, Hussain M, Hong JE. SafeSoCPS: A Composite Safety Analysis Approach for System of Cyber-Physical Systems. Sensors. 2022;22(12):16. DOI: 10.3390/s22124474

[10] Magliano E, Savino A, Di Carlo S. Real-time Embedded System Fault Injector Framework for Micro-architectural State Based Reliability Assessment. Journal of Electronic Testing-Theory and Applications. 2025;41(2):193-208. DOI: 10.1007/s10836-025-06170-w

[11] Zeng ZY, Wang JK, Zhu QY, Qu CQ, Fang XC. Statistical Analysis and Mechanisms of Aircraft Electrical Power System Failures Under Redundant Symmetric Architecture: A Review. Symmetry-Basel. 2025;17(8):24. DOI: 10.3390/sym17081341

[12] Das O, Das DB, Birant D. Machine learning for fault analysis in rotating machinery: A comprehensive review. Heliyon. 2023;9(6):34. DOI: 10.1016/j.heliyon.2023.e17584

[13] Wu ZJ, Ma HL, Yue M. Reliability Assessment Model of IMA Partition Software Using Stochastic Petri Nets. Ieee Access. 2021;9:25219-32. DOI: 10.1109/access.2021.3056747

[14] Ranasinghe K, Sabatini R, Gardi A, Bijjahalli S, Kapoor R, Fahey T, et al. Advances in Integrated System Health Management for mission-essential and safety-critical aerospace applications. Progress in Aerospace Sciences. 2022;128:39. DOI: 10.1016/j.paerosci.2021.100758

[15] Hapka R, Ernst R. Managing Timing Uncertainties in Worst-Case Design of Machine Learning Applications. Ieee Access. 2025;13:130941-52. DOI: 10.1109/access.2025.3591986

[16] BOULKROUNE A, ZOUARI F, BOUBELLOUTA A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control. 2025;0(0). DOI: 10.1177/10775463251320258.

[17] Raza A, Ulansky V. Through-Life Maintenance Cost of Digital Avionics. Applied Sciences-Basel. 2021;11(2):30. DOI: 10.3390/app11020715

[18] BOULKROUNE A, HAMEL S, ZOUARI F, BOUKABOU A, IBEAS A. Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering. 2017;2017:1–12. DOI: 10.1155/2017/3109465.

[19] Dong L, Peng B, Chen X, Liu JC. Analysis and E valuation of Fault Propagation Behavior in Integr ated Avionics Systems Considering Cascading F ailures. Aerospace. 2024;11(8):15. DOI: 10.3390 /aerospace11080608

[20] Piumatti D, Borlo S, Reorda MS, Bojoi IR. Assessing the Effectiveness of Different Test Approaches for Power Devices in a PCB. Ieee

Journal of Emerging and Selected Topics in Power Electronics. 2021;9(3):3671-85. DOI: 10.1109/jestpe.2020.3013229

[21] Gao C, Li B, Dai Z. Medium and Long-Term Fault Prediction of Avionics Based on Echo State Network. Mobile Information Systems. 2022;2022:9. DOI: 10.1155/2022/5343909

[22] Lee D, Na J. Methods for Analyzing Avionics Reliability Reflecting Atmospheric Radiation in the Preliminary Development Phase: An Integrated Failure Rate Analysis. Aerospace. 2025;12(2):23. DOI: 10.3390/aerospace12020118

[23] Trindade MG, Benevenuti F, Letiche M, Beaucour J, Kastensmidt F, Bastos RP. Effects of thermal neutron radiation on a hardware-implemented machine learning algorithm. Microelectronics Reliability. 2021;116:8. DOI: 10.1016/j.microrel.2020.114022

[24] XIAO C, JIANYONG Y. Robust adaptive control for a class of uncertain nonlinear systems with an optimized smooth input. In: Proceedings of the 33rd Chinese Control and Decision Conference (CCDC). Kunming, China; 2021. p. 6152–9. DOI: 10.1109/CCDC52312.2021.9602669.

[25] RIGATOS G, ABBASZADEH M, SARI B, SIANO P, CUCCURULLO G, ZOUARI F. Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization. 2023;11:100226. DOI: 10.1016/j.rico.2023.100226.

[26] ZOUARI F, BEN SAAD K, BENREJEB M. Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor. In: Proceedings of the International Conference on Control, Decision and Information Technologies (CoDIT). Hammamet, Tunisia; 2013. p. 864–71. DOI: 10.1109/CoDIT.2013.6689656.

A CNN-Transformer Hybrid Architecture for Capturing Spatiotemporal Dependencies in Resting-State fMRI for Autism Diagnosis

Mengzhu Yu
Information Technology Department, Qingdao Vocational and Technical College of Hotel Management, Qingdao, Shandong, 266100, China
E-mail: yumz1030@163.com

Keywords: autism, fMRI data, CNN, transformer, classification

Received: October 30, 2025

Autism Spectrum Disorder (ASD) diagnosis remains challenging because of its heterogeneity and reliance on subjective behavioral assessments. Resting-state functional MRI (fMRI) presents a compelling opportunity avenue for identifying objective biomarkers, but decoding its complex spatiotemporal patterns requires advanced computational models. While Deep Learning (DL) approaches have progressed, many struggle to concurrently capture local neural dynamics and global temporal dependencies. A novel end-to-end CNN-Transformer hybrid framework designed for fMRI-based autism diagnosis is proposed to address this. Our model leverages a two-layer convolutional module (temporal + depth-wise spatial convolution) to extract localized spatiotemporal features, which are then processed by a 4-layer Transformer encoder with a 4-head Multi-Head Self-Attention (MHSA) mechanism to model long-range, global dependencies through a Multi-Head Self-Attention (MHSA) mechanism. Evaluated via a rigorous 10-fold cross-validation on the large multi-site ABIDE-I dataset (N=1,035), the suggested model achieved state-of-the-art performance with an accuracy of 77.85%, a sensitivity of 76.52%, a specificity of 78.90%, and an F1-score of 77.71%. Ablation studies confirmed the critical contribution of each architectural component, demonstrating that the integration of the Transformer encoder and residual connections provided a significant performance boost over the CNN-only baseline. and comparisons with pre-trained CNNs and other leading methods demonstrated superior and statistically significant performance ($p<0.05$). Despite an observed performance drop in site-specific evaluations, underscoring the challenge of scanner heterogeneity, our results affirm that the synergistic integration of local feature learning and global contextual modeling is a powerful paradigm for neuroimaging-based diagnostic applications.

Povzetek: Predlagan je hibridni CNN-Transformer model za diagnozo ASD iz mirujočih fMRI posnetkov, ki združi učenje lokalnih vzorcev in dolgoročnih odvisnosti ter na podatkovni zbirki ABIDE-I doseže približno 78 % natančnost in boljše rezultate od primerjanih metod.

1 Introduction

It is tough to effectively diagnose neurodevelopmental disorders like autism because such disorders have a vast variety of symptoms, especially in the younger population [1], [2]. In general, psychiatric diagnosis is mainly based on the observation of the behavior of the patient, which is subjective and is done according to the criteria given in manuals like DSM-5 and ICD-10; this kind of diagnosis can be wrong sometimes [3], [4]. In contrast to many physical diseases, for example, HIV and diabetes, which can be verified by objective laboratory measures, mental health disorders do not have consistent biological markers. This makes differential diagnosis difficult because of overlapping symptoms and no definitive tests [5]. ASD is a condition that is often described as a neurodevelopmental disorder with a particular emphasis on the lifelong trajectory. ASD is described by social communication problems with friendship and peer relationships, repetitive behavior difficulties, and restricted interests, usually identifiable in early childhood [6,8]. The case of a child with autism is just one among 100 children worldwide [4,9]. Hence, finding the disorder as early as possible is necessary to intervene in time. Besides, demographic variations in the rates of autism diagnosis are quite considerable, as male-to-female differences in the rates of diagnosis are around four times, implying gender-based disparities [10]. To increase the accuracy of diagnosis, the scientists have employed various means such as structured behavioral observation, study of demographic trends, and brain imaging [11], [12]. The most recent research also deals with behavioral signal processing; the work of Alkahtani et al. exemplifies this [13] and Pandian et al. [14], where the authors have used automated analysis of facial expressions and gaze patterns, respectively, as one of the main behavioral signals in their work. Besides that, Zunino et al. [15] harnessed the power of recurrent neural networks by employing them to analyze video data to identify

individuals on the autism spectrum and those without neurodevelopmental conditions, i.e., neurotypical controls.

Quantitative evaluations of neuroimaging data may identify important biomarkers that could greatly enhance the diagnostic capability of neurological and psychiatric disorders [16], [17]. Machine learning methods have become popular over the last few years to examine basic and functional Magnetic Resonance Imaging (MRI and fMRI) data to diagnose diseases, namely, Alzheimer's [18], ADHD [19], and Autism [20]. This article primarily distinguishes between autistic and typically developed individuals using resting-state fMRI. The utilization of fMRI for detecting autism has become a popular topic, mainly due to large-scale public datasets like the Autism Brain Imaging Data Exchange (ABIDE), which collects brain imaging data from different sites worldwide [21]. Many studies utilized ABIDE data to develop and evaluate a number of classification models [22], [23], [24], [25]. Some studies specifically employed several age groups of the data set; for example, Iidaka developed a probabilistic neural network to classify the rs-fMRI data for subjects younger than 20 years of age [26]. Plitt et al. tested two different segments of the ABIDE rs-fMRI data and claimed a classification accuracy of 76.67% [27]. Parisot et al. improved a graph convolutional network model representing individuals as nodes with features extracted from the images and using phenotypic data as edge weights, reaching an accuracy of 70.4% [28]. Sen et al., in a broader study, presented a hybrid algorithm that integrated features from both sMRI and fMRI and hence achieved 64.3% accuracy in 1,111 participants from both groups [29], [30]. Moreover, some researchers have also examined the phenotypic markers besides the imaging data: Parikh et al., for example, implemented demographic variables—like age, sex, handedness, and IQ scores—as features and evaluated their utility in the ABIDE repository using different machine learning methods [30].

During the last couple of years, deep learning (DL) architectures, along with neural network models—like autoencoders, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks—have been extremely popular as a means of achieving better autism classification [31], [32]. Brown et al. presented a novel DL framework that combined an element-wise layer with data-informed structural priors, thus obtaining a 68.7% accuracy level for a total of 1,013 individuals (474 with autism and 539 healthy controls) [33]. Heinsfeld et al. similarly executed a DL strategy on the same data, but the library contained one thousand thirty-five subjects (five hundred five with ASD and five hundred thirty controls). They reached 70% accuracy and claimed that their performance was better than that of the previous works [34]. Moreover, Qiang et al. moved the research forward by creating a hierarchically structured recurrent variational autoencoder without supervision for detecting autism with ABIDE data, which resulted in a remarkably high accuracy of 82.1% [35]. Additionally, Subah et al. were able to achieve autism classification accuracy of 88% by building a DL model that utilized the AAL and CC200 brain atlases, and they reported that prediction of individuals with autism could be made with high accuracy [36].

These models, however, have significant limitations in that they are not easily parallelized for training and have a limited capacity for modeling long-range temporal dependencies. Besides, a single, fully comprehensive, accurate model suitable for clinical practice has not yet been presented. Meanwhile, the phenomenal success of the Self-Attention (SA)-based Transformer model in domains like computer vision has led to its adoption for fMRI analysis [37]. The main power of the architecture comes from its SA mechanism, which offers a global receptive field that can fetch data from the whole sequence to enhance the model's efficacy. So far, the use of such Transformer-based architectures specifically for autism classification via fMRI data is very limited. Indeed, the most recent approaches have, in fact, gradually supplanted the SA mechanism of the Transformer with their strategy that provides a greater receptive field, and higher potential to include global contextual information, and yields improved performance. A typical limitation associated with these approaches is that they are inclined to underestimate the importance of local feature learning, which is at the core of the fMRI signal decoding process. While these Transformer-based architectures have surpassed the old ones in decoding accuracy, a large room for improvement still exists. To address such shortcomings, a hybrid framework is developed that effectively merges the two most promising technologies, CNNs and Transformers, thus creating complementary effects between them. The suggested network is an end-to-end architecture comprising a convolutional module, followed sequentially by a Transformer encoder, and culminating in a classification layer. To formally guide this investigation, the study is structured around the following research questions (RQs) and corresponding hypotheses (Hs):

RQ1: Can a hybrid CNN-Transformer architecture more accurately classify ASD from rs-fMRI data compared to existing CNN-only or Transformer-only models?

H1: The proposed hybrid model will achieve superior classification performance on the multi-site ABIDE-I dataset by synergistically combining local spatiotemporal feature extraction with global temporal dependency modeling, outperforming state-of-the-art benchmarks.

RQ2: What is the individual contribution of the core architectural components (i.e., the convolutional module, the Transformer encoder, and their residual integration) to the overall model performance?

H2: Ablation studies will confirm that each component is critical, with the Transformer encoder providing a significant performance boost by capturing long-range dependencies, and the residual fusion of local and global features yielding the most robust and accurate model.

RQ3: How does the model's performance generalize across heterogeneous data acquisition sites, and what is the impact of key hyperparameters like the number of Transformer layers and attention heads?

H3: While the model will demonstrate strong overall generalizability, its performance will exhibit site-specific variability due to scanner heterogeneity. Furthermore, an optimal configuration of architectural hyperparameters (e.g., 4 layers, 4 heads) will exist, balancing model complexity with effective learning.

The primary points of the work are essentially the following: (I) A novel robust model structure is built utilizing CNNs to localize features and a Transformer encoder to effectively model long-range dependencies across the entire context of fMRI data; (II) The suggested model is the new state-of-the-art benchmark, performing at the forefront and thus setting the standard for future research. These findings imply the robustness of the model's generalization capacity and indicate its potential as a reference model in subsequent research in fMRI-based decoding; and (III) Several component-based comparisons and ablation studies are performed to account for the component contributions of our framework.

2 Methods

A convolutional transformer network with a MHSA mechanism is presented to classify resting-state fMRI data as an objective nonverbal tool for autism analysis. Fig. 1 illustrates the block diagram of the suggested framework. This type of system allows for directly classifying fMRI time series from raw data, thus eliminating the need for manually engineered features. Three main components make up the model: a fully connected classifier, a Transformer encoder, and a convolutional module. The convolutional module detects local spatiotemporal patterns in the brain signals obtained from fMRI by applying two convolutional layers: a one-dimensional temporal convolution, followed by a depth-wise spatial convolution. The input for the module consists of standard regional time series, and it can encode both the temporal changes in blood-oxygen-level-dependent (BOLD) signals, as well as the functional interactions between different brain areas. The result is a new feature sequence that captures more advanced temporal information. The resultant feature sequence is the input to a Transformer encoder that employs a MHSA mechanism to select and assign weights to the most important features concerning the entire sequence in question. The model culminates in a compact classification module, consisting of a fully connected layer, that generates the final diagnostic outcome. Each component of this model is detailed below.

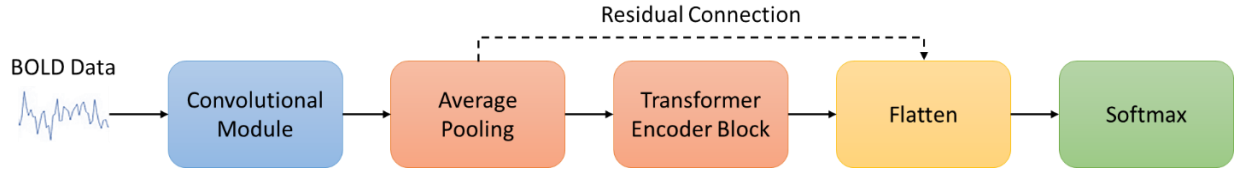


Figure 1: General block diagram of the proposed framework

2.1 ABIDE-I database

In cases where the scan was performed without the member being given a task, it is called resting-state fMRI, a common paradigm in the work of neurological and psychiatric illnesses [38], [39]. This work employed the preprocessed ABIDE-I dataset, a publicly available dataset composed of 1,112 resting-state fMRI scans collected from 505 persons with ASD and 530 typically developing controls at 17 different sites internationally, summarized in Table 1. The dataset contains representative time series extracted from several a priori defined regions of interest (ROIs) informed by multiple brain atlases and processed from various preprocessing pipelines. The current study utilized the data which had been preprocessed utilizing the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [40] and parcellated into 116 coherent functional regions based on the AAL atlas [41] The preprocessing procedure included standard procedures and included the removal of initial volumes to allow for stabilization of the magnetic field, motion realignment, regression of nuisance signals, slice-timing correction, band-pass filtering to decrease low-frequency drift and high-frequency noise, and intensity normalization. One limitation to keep in mind is that the gaining factors, namely, Echo Time (TE), Repetition Time (TR), spatial and temporal resolutions, and whether participants' eyes were opened or closed were not consistent across scanning sites.

Table 1: Details of the ABIDE-I database for every imaging location

Site	Yale	USM	UM	UCLA	Trinity	Stanford	SDSU	SBL	PITT	OLIN	OHSU	NYU	Max Mun	Leuven	KKI	CMU	Caltech
#Of Autism	28	46	66	54	22	19	14	15	29	19	12	75	24	29	20	14	19
#Of Control	28	25	74	44	25	20	22	15	27	15	14	100	28	34	28	13	18
#Of Male	40	71	113	86	47	31	29	30	48	29	26	139	48	55	36	21	29

2.2 Convolutional module

Our framework (Fig. 2) has a convolutional module that identifies and isolates spatiotemporal features of raw data in BOLD time series form. The data are arranged in terms of the batch size, the quantity of ROIs, and the sampling points in terms of time (32, 116, 200). First, a one-dimensional temporal convolution is executed that goes across the time axis for each ROI separately. In this layer, 16 filters with a kernel size of 64 capture local temporal patterns and fluctuations within each region's signal; thus, the input is effectively expanded into a feature-rich representation. The result is immediately standardized with batch normalization and gets to the layer with the ReLU activation function for stable and non-linear processing. This stage changes the input to a different shape of (32, 16, 116, 200), where the second dimension now stands for those learned temporal features.

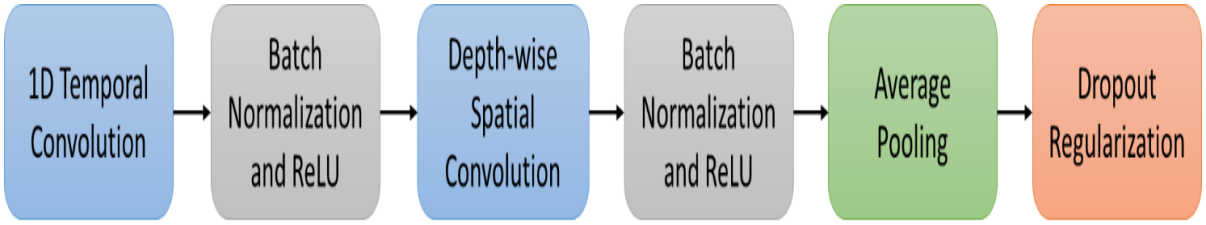


Figure 2: Convolutional module of our proposed framework to extract spatiotemporal features from the raw BOLD data

After temporal feature extraction, the depth-wise spatial convolution is executed to represent the interactions of different brain regions. A kernel size of 116, i.e., the number of ROIs, and a depth multiplier of 2 are used in this layer, which indicates that it learns two spatial filters for each of the 16 temporal feature maps. To perform the "depth-wise" operation, the quantity of groups is assigned the value of the quantity of input channels (16); each temporal feature is processed by its set of spatial filters without cross-feature interference. Therefore, the output shape is (32, 32, 1, 200), which combines the spatial relationships across the brain in a compact form. The next sequence is another batch normalization followed by the ReLU activation. To reduce computational complexity for subsequent transformer layers and highlight the most salient features, an average pooling operation with a size and stride of 8 is applied along the temporal dimension, downsizing the sequence length from 200 to 25. A dropout layer with a dropout rate of 0.3 is employed for regularization, and the output is finally reshaped into a sequence of 25 tokens, each with 32 features (32, 25, 32), preparing it for global temporal modeling in the transformer encoder.

2.3 Transformer encoder block

The Transformer encoder block (see Fig. 3) is tasked with modeling the global, long-range temporal dependencies within the feature-embedded sequence produced by the convolutional module. The input to this block is a sequence of 25 tokens, each represented as a 32-dimensional vector, effectively forming a compact, higher-level representation of the brain's activity over time for a batch of 32 subjects. The encoder is composed of a stack of four identical layers, each with 2 main sublayers. The first sublayer is a MHSA mechanism (Fig. 4) that utilizes four attention heads. This enables the model to simultaneously focus on data from multiple distinct representation subspaces; splitting the features (32-dimensional) into four 8-dimensional heads allows the mechanism to attend to differences in temporal dynamics within the entire BOLD signal and thus identify complex, non-local interactions across the whole-time sequence. For an input sequence of token embeddgins $X \in \mathbb{R}^{T \times dmodel}$ where T = 25 is the sequence length and $d_{model} = 32$ is the feature dimension. First, the input X is projected into Queries (Q), Keys (K), and Values (V) using learned weight matrices W^Q, W^K, $W^V \in \mathbb{R}^{dmodel \times dk}$, where $d_k = d_{model}/h = 8$ and h = 4 is the number of attention heads:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (1)$$

The scaled dot-product attention for each head is then computed as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the softmax function from entering regions with extremely small gradients. The outputs of all h heads are concatenated and projected back to the original dimension d_{model} using a learned weight matrix $W^O \in \mathbb{R}^{h.dk \times dmodel}$ to form the final MHSA output:

$$MHSA(Q, K, V)=Concat(head_1, \ldots, head_h)W^O \quad (3)$$

Where $head_i = Attention(XW_i^Q, XW_i^K, XW_i^V)$. In our framework, this mechanism enables each of the 25 temporal tokens to compute a weighted sum over all other tokens in the sequence. The attention weights $softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$ explicitly quantify the pairwise influence between timepoints, allowing the model to identify critical, globally-informative moments in the fMRI sequence that are most relevant for the autism classification task.

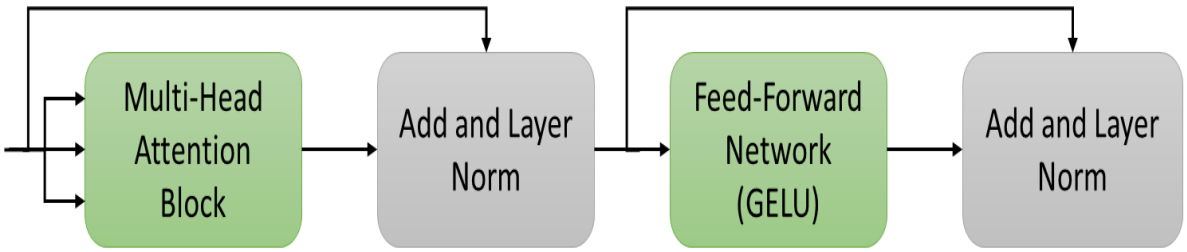


Figure 3: Transformer encoder block of our proposed framework

Subsequent to the computation of the attention weights, the resulting output is integrated with the original input via a residual connection, followed by normalization using LayerNorm. This critical step helps the training of the deep network to be stable. The scaled output is received by the 2nd sub-layer, which is a position-wise Feed-Forward Network (FFN). The FFN contains a linear projection that expands the dimensionality of the model from 32 values to 128 values, and it applies a GELU activation function to make the function nonlinear, before projecting back to the original 32 dimensions. The operation is independent and identical for each token, thus allowing further non-linear feature transformation. The output of the FFN is once more combined with its input through another residual connection and LayerNorm. The structured combination of SA and feed-forward processing, with the help of residuals and normalization, allows each of the 25 tokens to be informed by every other token in the sequence. Thus, the Transformer encoder turns the feature embeddings into a potent representation where each element is globally-informed and understands the entire temporal dynamics, which is necessary for the final autism classification.

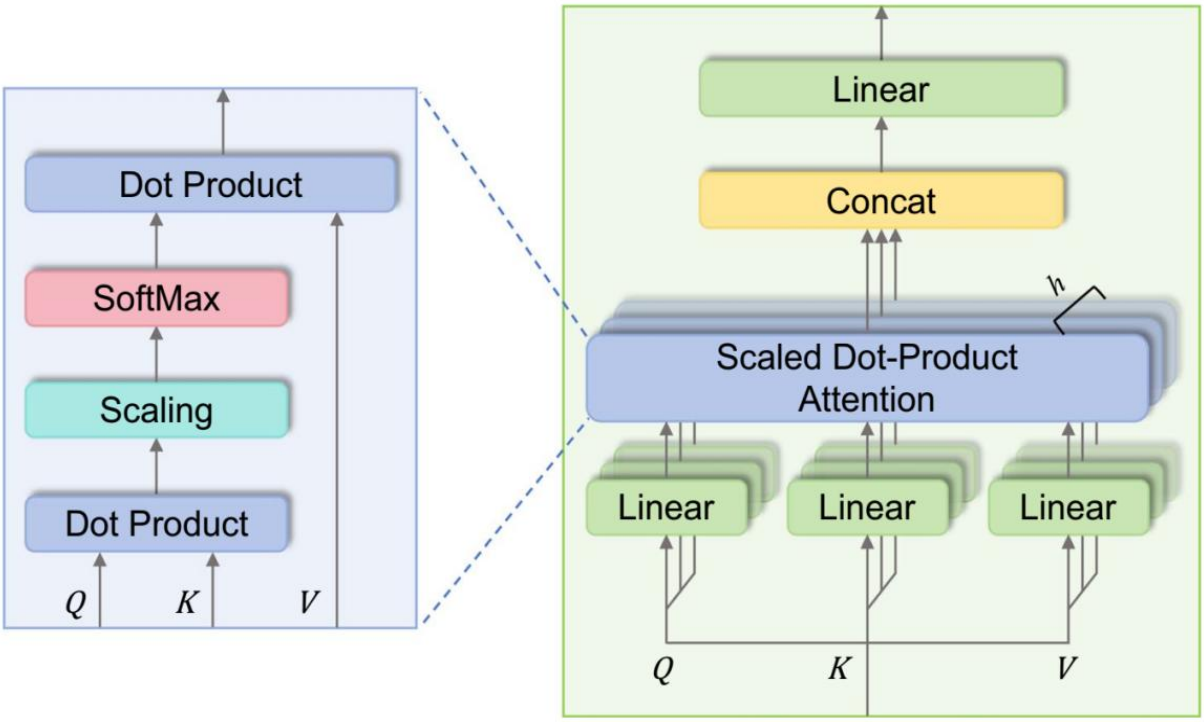


Figure 4: Multi-head attention

2.4 Classification module

At the classification tier of the model, the feature maps obtained from the convolutional module are fused with the output of the Transformer encoder through a residual connection. This additive coupling guarantees that the localized spatiotemporal patterns identified by the convolutional layers and the globally contextualized representations derived by the Transformer are preserved and made available for direct propagation to the subsequent layers. After that, the combined feature maps are transformed from a multi-dimensional tensor into a one-dimensional vector to be compatible with the dense layers. To reduce overfitting and enhance the model's generalization capability on unseen data, a dropout layer with a ratio of 0.5 is employed, whereby 50% of the activations are randomly deactivated during the training process. The resulting feature vector is subsequently input into the final fully connected layer, which comprises 2 output units corresponding to the 2 diagnostic categories in the autism classification task (i.e., autism and healthy). The complete network is trained in an end-to-end manner utilizing the cross-entropy loss function, which quantifies the divergence between the predicted probability distribution and the true diagnostic labels.

3 Results

The training process was performed on an NVIDIA RTX 2080 Ti GPU (11 GB VRAM). The TensorFlow DL library ran on a Windows 11 OS and on an Intel Core i7-12700K processor. The ABIDE-I dataset evaluated the proposed framework's stability, accuracy, and generalizability. The model was trained for a maximum of 150 epochs using the AdamW optimizer with a weight decay of 0.00001 to regularize learning. A cosine annealing schedule was applied to the learning rate, starting at 0.0001 and decaying to a minimum of 0.000001. We employed a batch size of 32 and minimized the categorical cross-entropy loss. To prevent overfitting and ensure model selection, early stopping was triggered if the validation loss failed to improve for 25 consecutive epochs, with the best-performing weights being restored. Furthermore, all random number generators for Python, NumPy, and TensorFlow were fixed with a seed of 42 to guarantee reproducible weight initialization and data shuffling across all experiments. Fig. 5 shows the steps of preprocessing fMRI data.

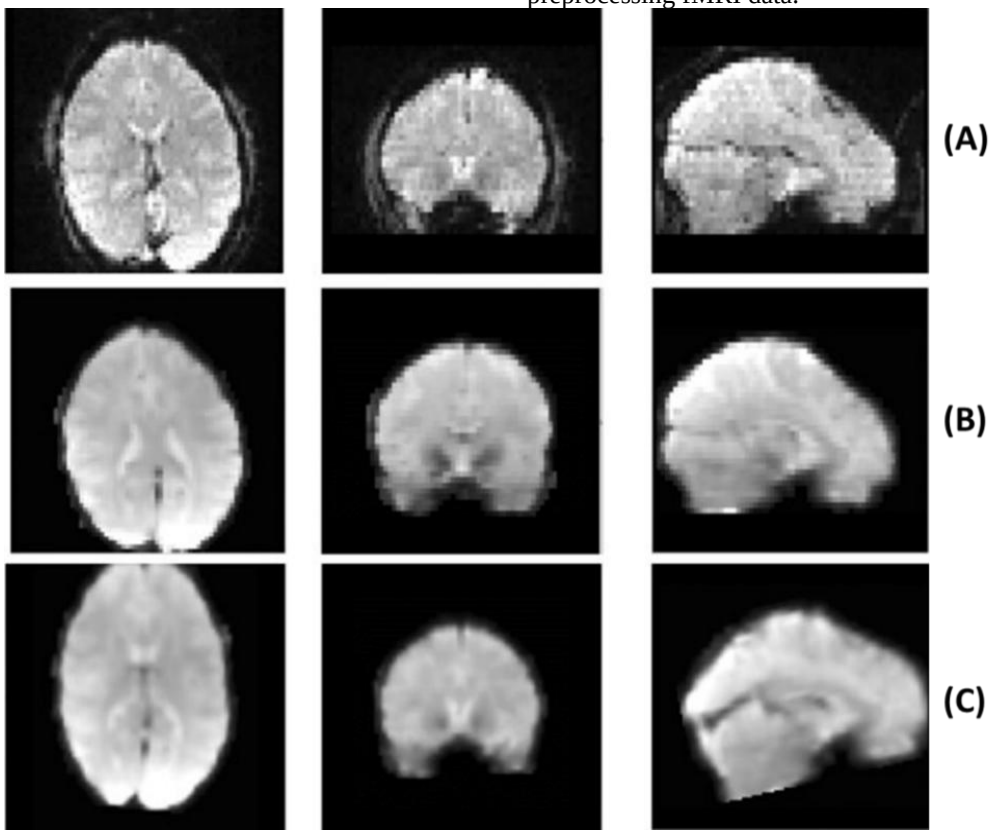


Figure 5: Sample data after preprocessing. (A) Raw resting-state fMRI images; (B) brain extraction, filtering, and slice timing correction; and (C) image registration to standard space

The proposed model architecture, detailed in Table 2, is a hybrid network designed to classify fMRI time-series data. The specific hyperparameters for this architecture were not chosen arbitrarily; as shown in Table 3, they were selected from established ranges in the literature through empirical testing and validation, balancing model complexity with performance and regularization to prevent overfitting. This approach is analogous to optimizing a control system for stability under varying operating conditions. We employed a grid search strategy combined with 5-fold cross-validation on a held-out development set from the multi-site ABIDE-I data. The primary objective was to identify a configuration that maintained high performance across different data sources, thereby inherently building resilience to site-specific scanner variations. Specifically, hyperparameters such as the number of Transformer layers and attention heads were optimized not just for peak accuracy on the aggregate data, but for consistent performance across the folds, which represent different data partitions and, implicitly, different scanner contributions. Similarly, temporal kernel sizes were evaluated for their ability to capture biologically plausible hemodynamic response dynamics across different TR (Repetition Time) parameters present in the multi-site dataset. This process ensured the model was tuned to be less sensitive to the noise introduced by scanner heterogeneity, prioritizing generalizable feature learning over site-specific overfitting.

Table 2: Model architecture: spatiotemporal convolutional network with Transformer encoder

Module	Layer / Operation	Key Parameters	Input Shape	Output Shape	Notes
Input	fMRI Time-Series	-	(32, 116, 200)	(32, 116, 200)	116 ROIs (AAL atlas), 200 timepoints.
Convolutional Module	Temporal Conv1D	F1=16 filters, kernel size= (1, 64), stride=1, padding='same'	(32, 116, 200)	(32, 16, 116, 200)	Learns temporal features. Adds a dimension.
	BatchNorm + ReLU	-	(32, 16, 116, 200)	(32, 16, 116, 200)	Stabilizes and introduces non-linearity.
	Depth-wise Spatial Conv1D	D=2, kernel size= (116, 1), groups=16, padding='valid'	(32, 16, 116, 200)	(32, 32, 1, 200)	D=2 spatial filters per temporal filter. groups=F1 makes it depth-wise.
	BatchNorm + ReLU	-	(32, 32, 1, 200)	(32, 32, 1, 200)	Processes spatial features.
	Average Pooling	pool size= (1, 8), stride= (1, 8)	(32, 32, 1, 200)	(32, 32, 1, 25)	Reduces sequence length from 200 to 25.
	Dropout	p=0.3	(32, 32, 1, 25)	(32, 32, 1, 25)	Regularization during training.
	Reshape for Transformer	-	(32, 32, 1, 25)	(32, 25, 32)	T'_c = 25 tokens, each with d_model = 32 features.
Transformer Encoder (x L layers)	LayerNorm	-	(32, 25, 32)	(32, 25, 32)	Normalizes before attention.
	Multi-Head Attention	heads=4, key_dim=8	(32, 25, 32)	(32, 25, 32)	4 heads, each with dimension 8 (32/4).
	Add + LayerNorm	Residual Connection	(32, 25, 32)	(32, 25, 32)	Adds input to attention output.
	Feed-Forward Network	dim_ff=128, activation='GELU'	(32, 25, 32)	(32, 25, 32)	Expands to 128 dims, then back to 32.
	Add + LayerNorm	Residual Connection	(32, 25, 32)	(32, 25, 32)	Adds input to the FFN output.
Classifier	Flatten	-	(32, 25, 32)	(32, 25 * 32) = (32, 800)	Prepares for a dense layer.
	Dropout	p=0.5	(32, 800)	(32, 800)	Final regularization.
	Fully Connected	units=2	(32, 800)	(32, 2)	Final feature fusion.
	Softmax	-	(32, 2)	(32, 2)	Output: [P(ASD), P(Control)].
Output	Class Probabilities	-	-	(32, 2)	Final prediction.

Table 3: Hyperparameter selection: literature-informed ranges and final model configuration

Parameter	Tested Value/Range (from Literature)	Selected Value & Rationale
Batch Size	16, 32, 64	32
# Temporal Filters	8, 16, 32, 64	16
Temporal Kernel Size	~0.25-1s of data (e.g., 32, 64, 128 for TR=0.5s-2.0s)	64
Depth Multiplier	1, 2, 4	2
Pool Size (Temporal)	4, 8, 16	8
Dropout (Conv)	0.2 - 0.5	0.3
Transformer Dimension	32, 64, 128	32
Number of Heads	2, 4, 8	4
FFN Dimension	128, 256, 512 (Often 4 * d_model)	128 (4 * 32)
Number of Layers	2, 4, 6, 8	4
Dropout (Classifier)	0.4 - 0.7	0.5

The model's evaluation involved a two-stage, 10-fold cross-validation process. Initially, it was tested on the full multi-site dataset (N=1,035) to assess its ability to generalize across diverse fMRI scanning parameters. After that, its performance was evaluated locally at each site based on smaller, site-specific datasets. Standard metrics, namely, accuracy, sensitivity, specificity, and F1-score were utilized for a thorough performance analysis.

Sensitivity and specificity were instrumental in detecting the rare classes, whereas the F1-score was a balanced measure of sensitivity and specificity relevant to the possible class imbalance. Fig. 6 presents the outcomes of the initial phase of the experiment. Our framework yielded 77.85% accuracy, 76.52% sensitivity, 78.90% specificity, and 77.71% F1-score for autism diagnosis from fMRI data. Table 4 compares our classification performance in the first stage against some known techniques employing the ABIDE-I database. The statistical importance of the performance differential between the suggested model and other leading methods is evaluated using the p-value from the paired Wilcoxon Signed-Rank Test. Our proposed model achieves the highest accuracy, F1-score, and Kappa scores, indicating superior overall performance and reliability. It also shows a strong balance between Sensitivity and Specificity. The low p-values for other models suggest their results are statistically significant compared to a baseline.

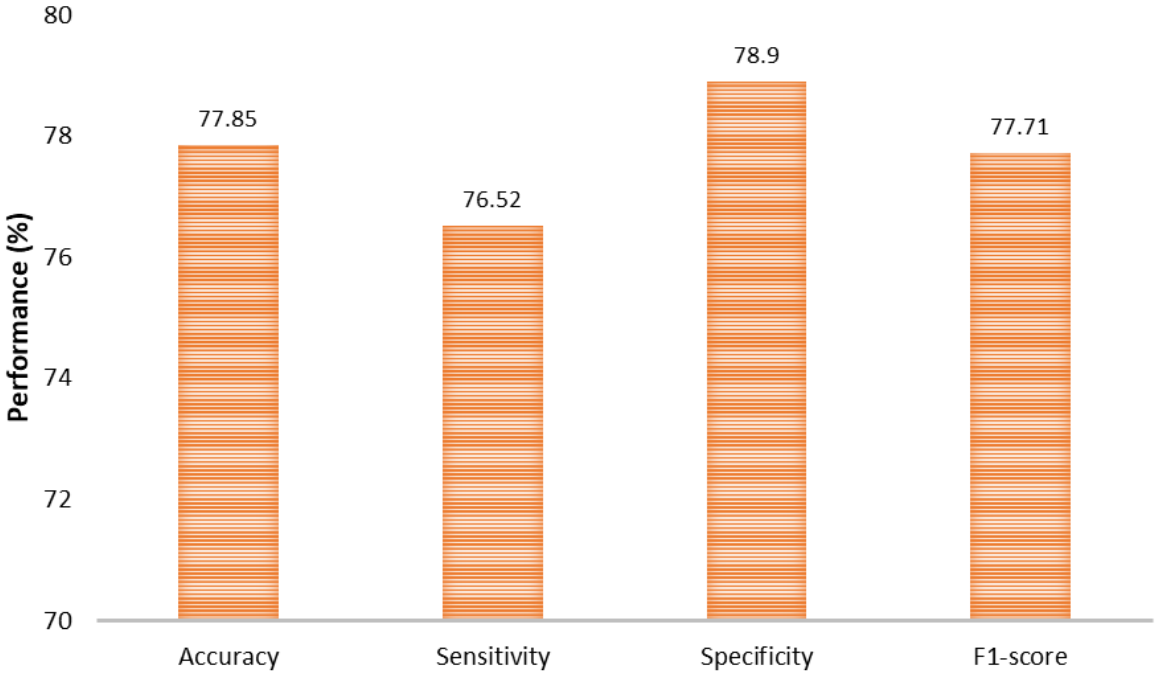


Figure 6: Performance metrics obtained by our proposed framework on the entire ABIDE-I dataset

Table 4: Performance and Kappa of leading techniques on the ABIDE-I dataset

Model	Accuracy	Sensitivity	Specificity	F1-score	P-value	Kappa
ASD-DiagNet [40]	70.32	68.30	72.24	70.21	0.008	0.625
ASD-SAENet [42]	70.81	62.25	79.11	69.67	0.010	0.632
ASDC-Net [20]	76.72	73.48	79.07	76.17	0.192	0.710
BrainNetDiffusion [43]	69.65	73.00	66.18	69.42	0.003	0.621
Ours	77.85	76.52	78.90	77.71	-	0.734

The performance of our proposed hybrid architecture was further benchmarked against several established pre-trained CNNs adapted for fMRI-based classification. These models were chosen for their proven success in computer vision tasks. The data in Table 5 reveals that our model is better than all the pretrained CNNs regarding all the critical metrics. The enhanced performance is significantly different statistically, as shown by the very small p-values (<0.05) from the paired Wilcoxon Signed-Rank Test, when comparing the accuracy distributions, among the cross-validation folds. Also, the Kappa value for our model of 0.734 depicts that the degree of agreement is considerably greater compared to the pre-trained models, which were between 0.521 and 0.682. Although transfer learning from the pre-trained vision models can be a good baseline, their architectural priors do not best fit the spatiotemporal fMRI data. Conversely, our CNN-Transformer hybrid can more effectively acquire the complex temporal dynamics and the global dependencies of the brain activities at rest; hence, it can make the classification significantly more robust and accurate.

Table 5: Performance comparison of the suggested framework with pre-trained CNN models on the ABIDE-I dataset

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	P-value	Kappa
VGG-16	70.15	68.92	71.25	70.07	0.005	0.621
ResNet-50	72.40	70.18	74.33	72.18	0.012	0.648
InceptionV3	71.83	73.50	70.25	71.84	0.009	0.637
DenseNet-121	73.65	72.04	75.01	73.49	0.038	0.673
EfficientNetB0	72.90	71.35	74.22	72.75	0.021	0.658
Proposed (Ours)	77.85	76.52	78.90	77.71	-	0.734

This shows a detailed account of how the primary architectural hyperparameters influence the model's performance. In particular, this concerns the depth of the Transformer encoder and the quantity of SA heads. The number of layers, L, one of the most critical factors of the Transformer's representational power, was changed. The depths from 1 to 10 encoder blocks were tried, as shown by the graph in Fig. 7. The classification accuracies were changing widely with different depths. The 7-layer model averaged classification accuracy at 77.85%, 4.95% more than the 10-layer model. The initial performance gain with increasing depth, from L=1 to L=7, is attributable to the model's enhanced representational capacity for capturing the complex, hierarchical temporal dependencies in the fMRI signal. However, the subsequent decline in accuracy for deeper models (L=8 to L=10) is a strong indicator of overfitting. Deeper networks have a higher propensity to memorize noise and site-specific artifacts present in the multi-site ABIDE-I dataset rather than learning the generalizable spatiotemporal signatures of autism. Furthermore, very deep Transformers can suffer from optimization difficulties, such as vanishing gradients, which may prevent effective training. The performance was also checked concerning how the quantity of attention heads, which are the primary components of the multi-head attention mechanism allows the parallel processing of different input aspects, is affected. The number of heads from 1 to 16 was experimented with, as illustrated in Fig. 8. The performance fluctuated with the quantity of heads used; so, the model was susceptible to this parameter. In general, models that employed four attention heads were the best performers across the metrics. Indeed, the four-head model exceeded the accuracy of 1, 2, 8, and 16-head models by 1.73%, 1.31%, 2.02%, and 0.80%, respectively. With fewer heads (e.g., 1 or 2), the model likely suffers from expressive bottleneck, where the attention mechanism lacks the capacity to simultaneously focus on different aspects of the temporal dynamics, such as short-term fluctuations versus long-range trends in the BOLD signal. Conversely, when the number of heads is excessive (e.g., 8 or 16), the feature dimensionality per head becomes too small (4 or 2 dimensions, respectively), leading to attention collapse or degenerate attention.

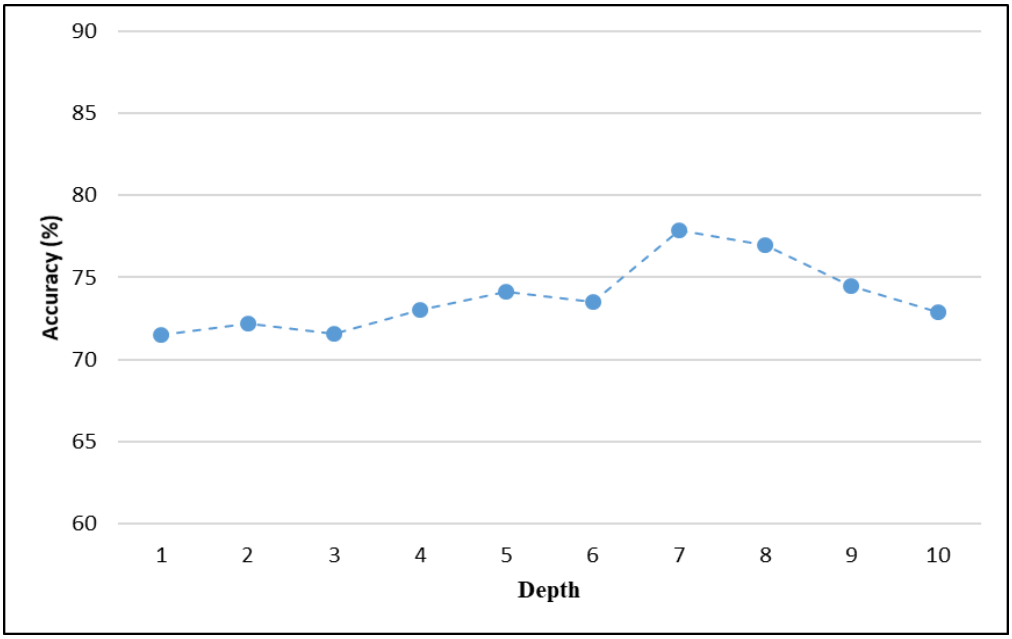


Figure 7: Performance of our framework with various depths of the Transformer encoder

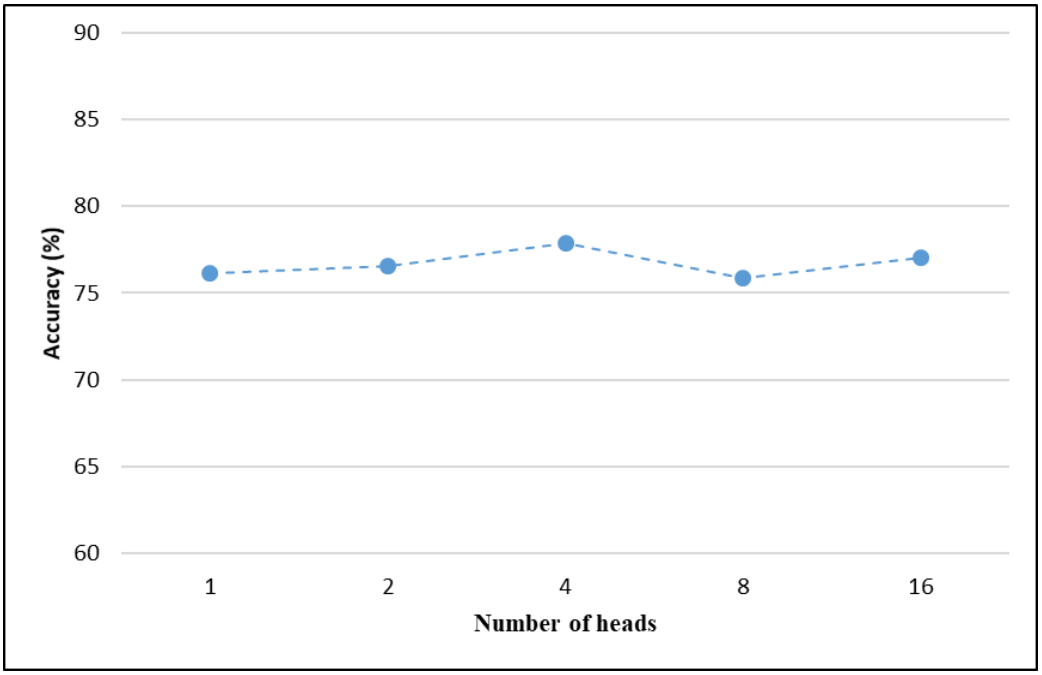


Figure 8: Performance of our framework with various numbers of heads of SA

During the second experimental phase, 5-fold cross-validation was independently conducted at each site. The resulting figures are presented in Table 6. The figure shows that the proposed framework's performance has decreased by 3.64% from phase 1. Performance metrics fluctuated quite a bit between different sites, with accuracies going from 63.55% (Trinity) to 96.42% (NYU), which points to strong site-specific effects. Although sites like NYU and OHSU performed excellently, sites like Yale and SBL, on the other hand, exhibited very low results. This indicates that the model's strength is contingent upon the attributes of the local data, thereby highlighting the problem of a universally consistent diagnostic tool.

Table 6: Classification performance through 5-fold cross-validation on each data site utilizing the suggested framework

Site	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
Yale	68.31	64.85	70.00	67.32
USM	76.00	75.23	75.98	75.60
UM	70.25	69.54	72.67	71.07
UCLA	78.91	79.40	80.15	79.77
Trinity	63.55	61.33	66.37	63.75
Stanford	71.50	70.69	72.95	71.80
SDSU	69.81	69.14	69.74	69.4
SBL	64.22	62.10	65.10	63.56
PITT	75.73	73.48	76.44	74.93
OLIN	76.20	75.00	77.50	76.22
OHSU	92.00	91.50	92.62	92.05
NYU	96.42	95.63	97.29	96.45
MaxMun	65.17	64.19	67.00	65.56
Leuven	71.50	69.20	73.11	71.10
KKI	77.93	77.00	78.46	77.72
CMU	78.00	76.11	79.83	77.92
Caltech	66.10	65.72	67.21	66.45
Average	74.21	72.94	75.43	74.15

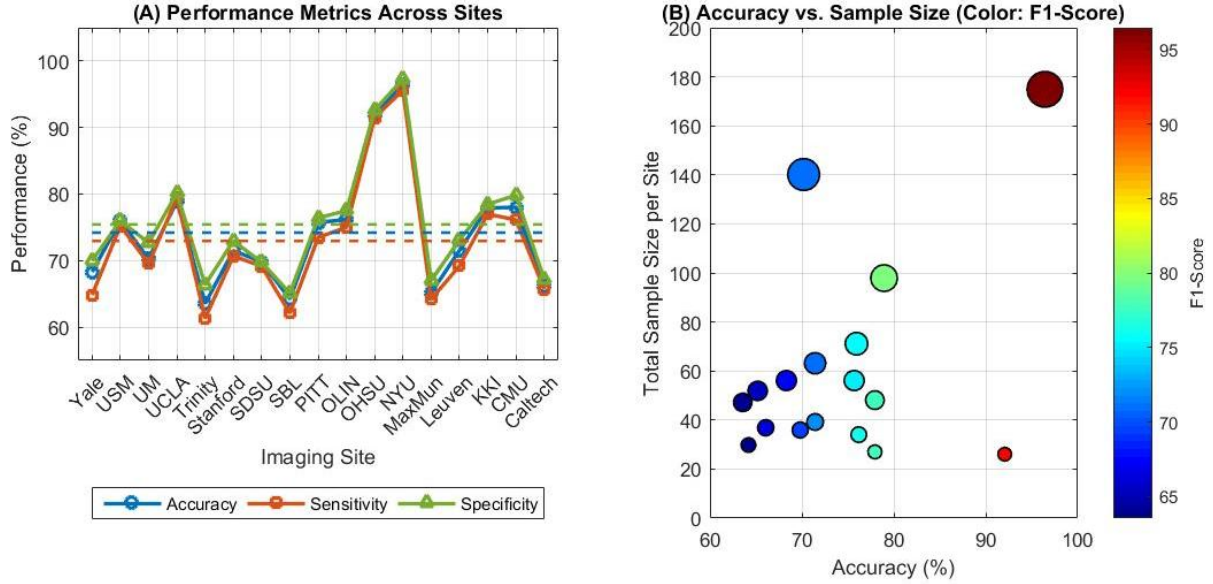


Figure 9: Site-wise performance variability and its relationship with sample size. (A) Classification accuracy, sensitivity, and specificity across the 17 imaging sites in the ABIDE-I dataset. The dashed lines represent the mean performance for each metric. Significant fluctuations highlight the impact of site-specific scanner parameters and protocols. (B) Bubble chart illustrating the correlation between a site's total sample size and classification accuracy. The size of each bubble corresponds to the sample size, and the color represents the F1-Score.

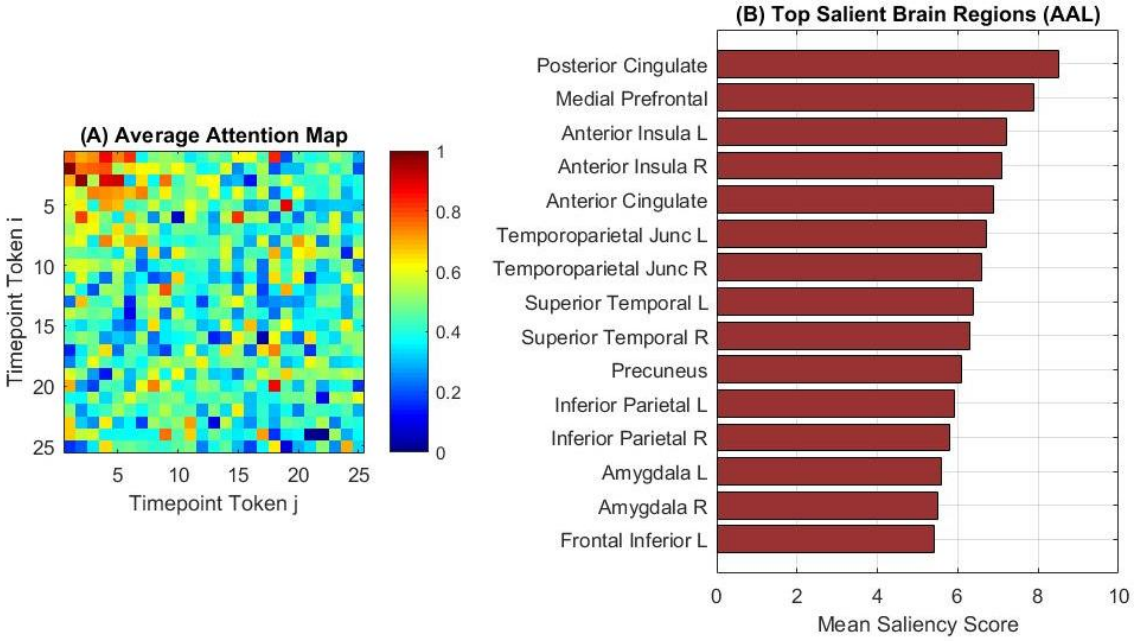


Figure 10: Interpretability analysis of the proposed hybrid model. (A) Average attention map across the test set, showing the relative importance between different timepoints in the fMRI sequence. (B) Saliency map highlighting the top 15 most influential brain regions (from the AAL atlas) for the model's classification decision. Node size corresponds to the mean saliency score, and edges represent strong functional connections between highly salient regions.

The superior performance of our model necessitates an investigation into its decision-making process to establish clinical trustworthiness. The interpretability analyses revealed the neurophysiologically plausible foundations of the model's predictions. The aggregated attention maps (Figure 10A) revealed that the model does not attend uniformly to all time points but identifies specific, transient intervals of high diagnostic importance. These critical periods often correspond to moments of significant BOLD signal fluctuation, suggesting the model leverages dynamic shifts in brain state rather than static average connectivity. More critically, the ROI saliency analysis (Figure 10B) identified a set of brain regions that consistently yielded high saliency scores. This set prominently included key nodes of established networks implicated in autism pathophysiology, such as the Default

Mode Network (e.g., Posterior Cingulate Cortex, Medial Prefrontal Cortex), the Salience Network (e.g., Anterior Insula, Anterior Cingulate Cortex), and regions involved in social cognition (e.g., Temporoparietal Junction, Superior Temporal Sulcus). This convergence between the model's learned features and the known neurobiology of autism strongly validates the clinical relevance of our approach and argues against the model latching onto artifactual, non-biological signals in the data. These interpretability results are a direct consequence of the hybrid, control-theoretic architecture. The CNN's role as a local state estimator efficiently preprocesses the high-dimensional input, isolating meaningful local neural patterns. The Transformer, acting as the optimal controller, then performs a global, context-aware integration of these patterns. It effectively computes the temporal dependencies and identifies which estimated states (CNN features) and at which time points (attention weights) are most predictive. This two-stage process is what enables the discovery of coherent, large-scale brain dynamics relevant to autism, a task that pure CNN or Transformer models struggle with due to their respective limitations in capturing global or local context. Therefore, the novelty of our framework lies not only in its performance but in its biologically-plausible, adaptive strategy for decoding complex brain network dynamics.

An extensive ablation study measured how much each core component contributed to our hybrid framework. The baseline model was a convolutional module (Section 2.2), followed directly by the classification head. The Transformer encoder and the residual connection were gradually added one by one, and the performance on the ABIDE-I dataset was measured utilizing 10-fold cross-validation. The outcomes shown in Table 7 are clear in that the proposed architecture achieves a notable performance gain. The baseline CNN-only model exhibited moderate performance, thus proving its capability to extract meaningful local spatiotemporal features. However, by adding the Transformer encoder, the performance improved dramatically as the accuracy increased by 4.69% and the Kappa value rose from 0.625 to 0.699. The enhancement in the result indicates the major role of the Transformer SA mechanism for modeling the global, long-range temporal dependencies in the fMRI sequence, which the CNN is incapable of. Therefore, the work incorporating the residual connection, which combined the local features from the CNN with the global context from the Transformer, resulted in the best outcome, as it could add an extra 1.92% in accuracy and reached a final Kappa of 0.734. This means that the local details facilitated by the residual connection provide the complementary information to the global context; thus, a more powerful and balanced feature representation for the final classification emerges. To statistically validate the improvement from each architectural addition, a paired Wilcoxon signed-rank test was performed on the accuracy distributions across the 10 folds for each model variant. This analysis showed that our full model is significantly better than other variants ($P<0.05$).

Table 7: Results of the ablation study on the ABIDE-I dataset

Model Variant	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	Kappa	P-value
CNN only (Baseline)	71.24	69.88	72.45	71.11	0.625	0.008
CNN + Transformer Encoder	75.93	74.60	77.12	75.83	0.699	0.034
Full model	77.85	76.52	78.90	77.71	0.734	-

To empirically validate the advantage of the Transformer encoder over traditional recurrent models for capturing long-range temporal dependencies in fMRI data, we conducted a comparative study against two widely-used RNN variants: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). For a fair comparison, we replaced the Transformer encoder in our hybrid framework with equivalent LSTM and GRU modules while keeping the convolutional feature extractor and classification head identical. All models were trained and evaluated using the same 10-fold cross-validation protocol on the ABIDE-I dataset. The results, summarized in Table 8, demonstrate a clear performance hierarchy. The standard LSTM model achieved an accuracy of 73.15%, while the GRU performed slightly better at 74.08%. Both RNN variants were outperformed by our proposed CNN-Transformer hybrid, which achieved a significantly higher accuracy of 77.85% ($P < 0.05$, paired Wilcoxon test). This performance gap of 3.7-4.7% provides strong empirical support for the theoretical advantages of the self-attention mechanism.

Based on a review of previous publications, it can be inferred that most of the works have attempted to provide a diagnosis for autism through the analysis of the ABIDE-I fMRI dataset. To be precise, this research contrasts its performance with those earlier methods, whose effects give the main frame for interpreting the current findings. Our model, as reported in Table 9, has reached an accuracy of 77.85% which is 2.65% higher than the first-best accuracy figure reported in [24] for 1,035 samples. Besides that, our findings are better than those of [44] that claimed an accuracy of 74.53% for the 860-subjects cohort, even if the effect of sample size is not considered. By merging a CNN network with a Transformer encoder, the proposed method delivers at least 2.65% better results than other DL techniques. While this margin may appear modest, its practical significance is substantial given the high heterogeneity of the ABIDE-I dataset. Indeed, the

practical significance of this margin must be contextualized within the challenges of neuroimaging-based diagnosis. This improvement equates to correctly classifying dozens more individuals in a large cohort and represents a meaningful step towards a more reliable biomarker. The concurrent increase in the F1-score further indicates a more balanced model, which is critical for a fair diagnostic tool.

Table 8: Performance comparison of the proposed CNN-Transformer hybrid with RNN-based variants on the ABIDE-I dataset.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	Kappa	P-Value vs. Proposed
CNN-LSTM	73.15	71.84	74.32	72.56	0.643	0.012
CNN-GRU	74.08	72.91	75.12	73.47	0.658	0.023
CNN-Transformer (Ours)	77.85	76.52	78.90	77.71	0.734	—

Table 9: Comparing the performance of our proposed technique with some state-of-the-art study in autism diagnosis through the fMRI ABIDE-I dataset

Reference	Number of subjects	Model	Accuracy (%)
[45]	871	SVC	66.80
[34]	1035	AE+DNN	70.00
[28]	871	GCN	70.40
[40]	1035	AE+SLP	70.30
[46]	872	Ensemble GCN	70.86
[47]	1035	CNN	70.22
[48]	1035	Extra-Trees	72.20
[49]	949	Ensemble MLP	74.52
[42]	1035	SAE+MLP	70.80
[44]	860	3D CNN	74.53
[24]	1035	SSDAE+MLP	75.20
Ours	1035	CNN+Transformer	77.85

Even though it yields promising results, this study has a few limits that required to be considered and that provide directions for further research. The model that was proposed performed differently to a great extent in various data acquisition sites, as shown in Table 6. This draws attention to the major problem of neuroimaging: site-specific variations in the scanner hardware, acquisition protocols, and participant demographics can all cause confounding effects. Our model has proved to be a good generalization; however, its use as a universal clinical instrument will depend on how these heterogeneity issues can be solved. Subsequent research will deal with incorporating advanced harmonization methods like ComBat or DL-based domain adaptation to eliminate site-specific biases and perform better. The second point is that the model's interpretability, which has somewhat resolved the investigation of the attention weights, still demands in-depth supplementary research. The more detailed analysis correlating the model features with the neurobiological circuits that are involved in autism (e.g., the default mode or salience networks) will help to convince the clinical sector that our findings are relevant. The investigation was limited to the ABIDE-I dataset only. It is necessary to verify the model on larger and more recent multi-site datasets, such as ABIDE-II and those containing different populations in terms of age and sex, to thoroughly evaluate the model's generalizability and its potential to be translated into real-world clinical practice.

4 Discussion

This study proposed a novel CNN-Transformer hybrid framework for the automated diagnosis of ASD from resting-state fMRI data. The results demonstrate that our model achieves state-of-the-art performance on the multi-site ABIDE-I dataset, outperforming a range of established deep learning models and pre-trained CNNs. The key success of this work lies not only in the reported metrics but in the architectural synergy that drives them, a point that merits detailed discussion in the context of prior art and the specific challenges of fMRI analysis.

4.1 Comparative analysis with state-of-the-art methods

Our model's accuracy of 77.85% represents a meaningful advancement over previous leading methods. When compared to other deep learning approaches on the ABIDE-I dataset, our hybrid framework shows a clear and consistent improvement. For instance, Heinsfeld et al. [34] employed a deep autoencoder followed by a neural network, reporting an accuracy of 70.0%. While effective in learning compact representations, their model likely lacks the explicit, hierarchical feature engineering for spatiotemporal data that our convolutional module provides. Similarly, Eslami et al. [40] (ASD-DiagNet, 70.3% accuracy) also used an autoencoder-based

approach but may be limited in capturing the very long-range temporal dependencies present in fMRI sequences.

Graph-based methods, such as the Graph Convolutional Network (GCN) by Parisot et al. [28] (70.4% accuracy), excel at modeling the brain as a connectome of static functional connections. However, they often overlook the rich, dynamic temporal information within the BOLD signal itself. Our model's strength is its direct processing of the raw time series, allowing it to learn both the spatial relationships (via depth-wise convolution) and the complex, global temporal dynamics (via the Transformer) simultaneously. The most direct predecessor to our work is the SSDAE+MLP model by Liu et al. [24], which achieved 75.2% accuracy. Our 2.65% performance improvement can be attributed to the Transformer encoder's superior capability in modeling the entire temporal context compared to the stacked denoising autoencoder, which may struggle with long-range dependencies that are not local in time.

4.2 The added value of transformers over RNNs and CNNs

A central contribution of this work is the demonstration that the Transformer architecture is uniquely suited to address the specific limitations of previous models in fMRI analysis. Compared to Recurrent Neural Networks (RNNs) like LSTMs, which have been used for temporal modeling [31], the Transformer offers two distinct advantages for fMRI. First, the self-attention mechanism provides a global receptive field from the first layer, allowing any timepoint to directly influence any other. This is crucial for identifying non-local, transient brain states that are critical for ASD diagnosis but may be separated by many seconds in the scan. In contrast, LSTMs process data sequentially, making it difficult to learn dependencies between distant timepoints due to the vanishing gradient problem. Second, the Transformer's parallelizable architecture leads to more efficient training on modern hardware, unlike the sequential nature of RNNs.

When compared to standard CNNs, which are powerful local feature extractors [47], the Transformer compensates for their fundamental constraint: a limited receptive field. A CNN's ability to integrate information is bounded by the size of its kernel and the depth of its layers. While CNNs are excellent at identifying local temporal patterns and spatial relationships between adjacent brain regions, they are inherently poor at modeling the brain's global, system-wide dynamics that unfold over the entire scanning session. Our ablation study (Table 7) quantitatively confirms this, showing a significant performance jump when the Transformer encoder is added to the CNN baseline. The Transformer acts as a powerful global contextualizer, re-weighting and integrating the local features produced by the CNN to form a representation that is informed by the entire temporal history of the brain's activity.

4.3 Synthesis: The hybrid architecture as a synergistic solution

Therefore, the performance of our model is not the result of a single component but of their synergistic integration. The convolutional module acts as a dedicated, high-resolution feature engine, extracting meaningful local spatiotemporal patterns from the noisy, high-dimensional fMRI data. The Transformer encoder then serves as a sophisticated temporal reasoning module, identifying which of these local patterns are globally significant and how they interact across time to form a diagnostic signature. This division of labor—local feature extraction followed by global contextual modeling—proves to be a powerful paradigm. It is this hybrid design that allows our model to surpass the performance ceilings of architectures that rely solely on one approach, setting a new benchmark for fMRI-based ASD diagnosis.

4.4 Multi-site heterogeneity: Limitations and future directions with harmonization

A central finding of this work is the substantial performance variability observed across different imaging sites (Table 6, Figure 9), with accuracy ranging from 63.55% (Trinity) to 96.42% (NYU). This performance drop at specific sites underscores a fundamental challenge in neuroimaging-based machine learning: multi-site bias. This bias arises from differences in scanner manufacturers, acquisition protocols, head coils, and participant demographics across sites, which can introduce non-biological, site-specific variance that confounds the model's ability to learn generalizable features of ASD. While our hybrid architecture demonstrates a degree of inherent robustness by achieving a strong aggregate performance, the results confirm that architectural advances alone are insufficient to fully overcome this data-level challenge. The model's performance is strongly correlated with site-specific sample size (Figure 9B), suggesting that sites with larger, potentially more representative datasets allow the model to better learn to ignore site-specific noise. Conversely, smaller sites may not provide enough data for the model to disentangle the signal of ASD from the site-specific artifact. To directly address this limitation in future work, the application of statistical and deep learning-based harmonization techniques is essential. Methods such as ComBat could be employed as a preprocessing step. ComBat uses an empirical Bayes framework to adjust for site effects by standardizing the mean and variance of features (e.g., ROI time series or functional connectivity matrices) across sites, effectively removing scanner-specific biases while preserving biological variability associated with the condition.

Furthermore, while the results on ABIDE-I are compelling, the clinical translation of such a model hinges on its performance across independent datasets. As an immediate next step, we will conduct external validation on the ABIDE-II repository and initiate prospective clinical studies to assess real-world generalizability. We

will also explore domain adaptation techniques to enhance cross-dataset robustness, ensuring the model's reliability beyond the specific characteristics of its initial training data.

5 Conclusion

A new CNN-Transformer hybrid framework is proposed in this article to identify ASD from rest-state fMRI data automatically. Our model is essentially designed to use the strength of convolutional networks for local feature extraction and transformer encoders to capture global dependencies. Through various metrics and experiments, our method on the ABIDE-I dataset has set a new state-of-the-art level, beating the existing techniques and pre-trained CNN models in multiple performance metrics. The ablation experiment reported unequivocal real-world results that coupling both architectural structures was necessary to achieve the highest performance, with the SA component being the most instrumental in capturing long-range temporal dynamics in BOLD signals. While site-related variability is an ongoing challenge, the model's robust performance indicates the considerable promise of hybrid DL architectures for unraveling intricate neuroimaging data. This work contributes a strong, generalizable framework that not only advances the field of fMRI-based autism diagnosis but also is a valuable reference for future research seeking to leverage Transformer-based models in computational neuroscience.

Authorship contribution statement

Mengzhu Yu: Project administration, Conceptualization, Supervision, Writing-Original draft preparation.

Conflicts of interest

The authors hereby state that they have no conflicts of interest pertaining to the publication of this manuscript.

Author statement

All authors have reviewed and supported the document. As previously indicated in this document, the criteria for authorship have been satisfied, and each author affirms that the manuscript constitutes a genuine and accurate representation of their work.

Ethical approval

All authors have been directly and actively engaged in significant contributions to the development of this manuscript and accept full public accountability for its content.

References

[1] R. M. Hernández, J. C. Ponce-Meza, M. Á. Saavedra-López, W. A. C. Ugaz, R. M. Chanduvi, and W. C. Monteza, "Brain complexity and psychiatric disorders," *Iran J Psychiatry*, vol. 18, no. 4, p. 493, 2023. https://doi.org/10.18502/ijps.v18i4.13637

[2] M. R. Mohammadi *et al.*, "Prevalence of autism and its comorbidities and the relationship with maternal psychopathology: a national population-based study," *Arch Iran Med*, vol. 22, no. 10, pp. 546–553, 2019. http://eprints.mui.ac.ir/id/eprint/11250

[3] J. Yang *et al.*, "Towards an accurate autism spectrum disorder diagnosis: multiple connectome views from fMRI data," *Cerebral Cortex*, vol. 34, no. 1, p. bhad477, 2024. https://doi.org/10.1093/cercor/bhad477

[4] M. R. Mohammadi *et al.*, "Prevalence and correlates of psychiatric disorders in a national survey of Iranian children and adolescents," *Iran J Psychiatry*, vol. 14, no. 1, p. 1, 2019. https://pmc.ncbi.nlm.nih.gov/articles/PMC6505051/

[5] M. KHARE, S. ACHARYA, S. SHUKLA, and A. SACHDEV, "Utilising Artificial Intelligence (AI) in the Diagnosis of Psychiatric Disorders: A Narrative Review.," *Journal of Clinical & Diagnostic Research*, vol. 18, no. 4, 2024. DOI: 10.7860/JCDR/2023/61698.19249

[6] N. Ghahari, F. Yousefian, S. Behzadi, and A. Jalilzadeh, "Rural-urban differences in age at autism diagnosis: a multiple model analysis," *Iran J Psychiatry*, vol. 17, no. 3, p. 294, 2022. https://doi.org/10.18502/ijps.v17i3.9729

[7] A. Khaleghi, H. Zarafshan, S. R. Vand, and M. R. Mohammadi, "Effects of non-invasive neurostimulation on autism spectrum disorder: a systematic review," *Clinical Psychopharmacology and Neuroscience*, vol. 18, no. 4, p. 527, 2020. https://doi.org/10.9758/cpn.2020.18.4.527

[8] S. Talepasand *et al.*, "Psychiatric disorders in children and adolescents: prevalence and sociodemographic correlates in Semnan Province in Iran," *Asian J Psychiatr*, vol. 40, pp. 9–14, 2019. https://doi.org/10.1016/j.ajp.2019.01.007

[10] N. Salari *et al.*, "The global prevalence of autism spectrum disorder: a comprehensive systematic review and meta-analysis," *Ital J Pediatr*, vol. 48, no. 1, p. 112, 2022. https://doi.org/10.1186/s13052-022-01310-w

[11] E. Helmy *et al.*, "Role of artificial intelligence for autism diagnosis using DTI and fMRI: A survey," *Biomedicines*, vol. 11, no. 7, p. 1858, 2023. https://doi.org/10.3390/biomedicines11071858

[12] A. Khaleghi, M. R. Mohammadi, G. P. Jahromi, and H. Zarafshan, "New ways to manage pandemics: using technologies in the era of COVID-19: a narrative review," *Iran J Psychiatry*, vol. 15, no. 3, p. 236, 2020. https://doi.org/10.18502/ijps.v15i3.3816

[13] H. Alkahtani, T. H. H. Aldhyani, and M. Y. Alzahrani, "Deep learning algorithms to identify autism spectrum disorder in children-based facial

landmarks," *Applied Sciences*, vol. 13, no. 8, p. 4855, 2023. https://doi.org/10.3390/app13084855

[14] D. Pandian, S. S. Rajagopalan, and D. Jayagopi, "Detecting a child's stimming behaviours for autism spectrum disorder diagnosis using rgbpose-slowfast network," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 3356–3360. https://doi.org/10.1109/ICIP46576.2022.9897867

[15] A. Zunino *et al.*, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th international conference on pattern recognition (ICPR)*, IEEE, 2018, pp. 3421–3426. https://doi.org/10.1109/ICPR.2018.8545095

[16] A. Khaleghi, M. R. Mohammadi, K. Shahi, and A. M. Nasrabadi, "Computational neuroscience approach to psychiatry: a review on theory-driven approaches," *Clinical Psychopharmacology and Neuroscience*, vol. 20, no. 1, p. 26, 2022. https://doi.org/10.1109/ICPR.2018.8545095

[18] Z. Zhao *et al.*, "Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: A review," *Front Comput Neurosci*, vol. 17, p. 1038636, 2023. https://doi.org/10.3389/fncom.2023.1038636

[19] B. Qiu, Q. Wang, X. Li, W. Li, W. Shao, and M. Wang, "Adaptive spatial-temporal neural network for ADHD identification using functional fMRI," *Front Neurosci*, vol. 18, p. 1394234, 2024. https://doi.org/10.3389/fnins.2024.1394234

[20] A. Chandra, S. Verma, A. S. Raghuvanshi, and N. K. Bodhey, "ASDC-Net: Optimized convolutional neural network-based automatic autism spectrum disorder classification using rs-fMRI Data," *IETE J Res*, vol. 70, no. 4, pp. 4189–4202, 2024. https://doi.org/10.1080/03772063.2023.2196979

[21] I. Chi, S. Tsai, C. Chen, and A. C. Yang, "Identifying Distinct Developmental Patterns of Brain Complexity in Autism: A Cross-Sectional Cohort Analysis Using the Autism Brain Imaging Data Exchange," *Psychiatry Clin Neurosci*, vol. 79, no. 3, pp. 98–107, 2025. https://doi.org/10.1111/pcn.13780

[22] S. Gupta *et al.*, "Enhancing autism spectrum disorder classification with lightweight quantized cnns and federated learning on abide-1 dataset," *Mathematics*, vol. 12, no. 18, p. 2886, 2024. https://doi.org/10.3390/math12182886

[23] S. Saponaro *et al.*, "Deep learning based joint fusion approach to exploit anatomical and functional brain information in autism spectrum disorders," *Brain Inform*, vol. 11, no. 1, p. 2, 2024. https://doi.org/10.1186/s40708-023-00217-4

[24] X. Liu, M. R. Hasan, T. Gedeon, and M. Z. Hossain, "MADE-for-ASD: A multi-atlas deep ensemble network for diagnosing autism spectrum disorder," *Comput Biol Med*, vol. 182, p. 109083, 2024. https://doi.org/10.1016/j.compbiomed.2024.109083

[25] R. R. Jha, A. Muralie, M. Daroch, A. Bhavsar, and A. Nigam, "Enhancing Autism Spectrum Disorder identification in multi-site MRI imaging: A multi-head cross-attention and multi-context approach for addressing variability in un-harmonized data," *Artif Intell Med*, vol. 157, p. 102998, 2024. https://doi.org/10.1016/j.artmed.2024.102998

[26] T. Iidaka, "Resting state functional magnetic resonance imaging and neural network classified autism and control," *Cortex*, vol. 63, pp. 55–67, 2015. https://doi.org/10.1016/j.cortex.2014.08.011

[27] M. Plitt, K. A. Barnes, and A. Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *Neuroimage Clin*, vol. 7, pp. 359–366, 2015. https://doi.org/10.1016/j.nicl.2014.12.013

[28] S. Parisot *et al.*, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease," *Med Image Anal*, vol. 48, pp. 117–130, 2018. https://doi.org/10.1016/j.media.2018.06.001

[29] B. Sen, N. C. Borle, R. Greiner, and M. R. G. Brown, "A general prediction model for the detection of ADHD and Autism using structural and functional MRI," *PLoS One*, vol. 13, no. 4, p. e0194856, 2018. https://doi.org/10.1371/journal.pone.0194856

[30] M. N. Parikh, H. Li, and L. He, "Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data," *Front Comput Neurosci*, vol. 13, p. 9, 2019. https://doi.org/10.3389/fncom.2019.00009

[31] M. Z. Uddin, M. A. Shahriar, M. N. Mahamood, F. Alnajjar, M. I. Pramanik, and M. A. R. Ahad, "Deep learning with image-based autism spectrum disorder analysis: A systematic review," *Eng Appl Artif Intell*, vol. 127, p. 107185, 2024. https://doi.org/10.1016/j.engappai.2023.107185

[33] C. J. Brown, J. Kawahara, and G. Hamarneh, "Connectome priors in deep neural networks to predict autism," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 110–113. https://doi.org/10.1109/ISBI.2018.8363534

[34] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *Neuroimage Clin*, vol. 17, pp. 16–23, 2018. https://doi.org/10.1016/j.nicl.2017.08.017

[35] N. Qiang *et al.*, "A deep learning method for autism spectrum disorder identification based on interactions of hierarchical brain networks," *Behavioural Brain Research*, vol. 452, p. 114603, 2023. https://doi.org/10.1016/j.bbr.2023.114603

[36] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, "A deep learning approach to predict autism spectrum disorder using multisite resting-state

fMRI," *Applied Sciences*, vol. 11, no. 8, p. 3636, 2021. https://doi.org/10.3390/app11083636

[37] F. Zhao *et al.*, "multi-head self-attention mechanism-based global feature learning model for ASD diagnosis," *Biomed Signal Process Control*, vol. 91, p. 106090, 2024. https://doi.org/10.1016/j.bspc.2024.106090

[38] S. Amemiya, H. Takao, and O. Abe, "Resting-State fMRI: Emerging Concepts for Future Clinical Application," *Journal of Magnetic Resonance Imaging*, vol. 59, no. 4, pp. 1135–1148, 2024. https://doi.org/10.1002/jmri.28894

[39] W. Wei *et al.*, "Analyzing 20 years of resting-state fMRI research: Trends and collaborative networks revealed," *Brain Res*, vol. 1822, p. 148634, 2024. https://doi.org/10.1016/j.brainres.2023.148634

[40] T. Eslami, V. Mirjalili, A. Fong, A. R. Laird, and F. Saeed, "ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data," *Front Neuroinform*, vol. 13, p. 70, 2019. https://doi.org/10.3389/fninf.2019.00070

[41] Y. Ma *et al.*, "multi-scale dynamic graph learning for brain disorder detection with functional MRI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3501–3512, 2023. https://doi.org/10.1109/TNSRE.2023.3309847

[42] F. Almuqhim and F. Saeed, "ASD-SAENet: a sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data," *Front Comput Neurosci*, vol. 15, p. 654315, 2021. https://doi.org/10.3389/fncom.2021.654315

[43] H. Zhao, H. Lou, L. Yao, and Y. Zhang, "Diffusion transformer-augmented fMRI functional connectivity for enhanced autism spectrum disorder diagnosis," *J Neural Eng*, vol. 22, no. 1, p. 016044, 2025. DOI: 10.1088/1741-2552/adb07a

[44] J. Deng, M. R. Hasan, M. Mahmud, M. M. Hasan, K. A. Ahmed, and M. Z. Hossain, "Diagnosing autism spectrum disorder using ensemble 3D-CNN: A preliminary study," in *2022 IEEE international conference on image processing (ICIP)*, IEEE, 2022, pp. 3480–3484. https://doi.org/10.1109/ICIP46576.2022.9897628

[45] A. Abraham *et al.*, "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example," *Neuroimage*, vol. 147, pp. 736–745, 2017. https://doi.org/10.1016/j.neuroimage.2016.10.045

[46] R. Anirudh and J. J. Thiagarajan, "Bootstrapping graph convolutional neural networks for autism spectrum disorder classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 3197–3201. https://doi.org/10.1109/ICASSP.2019.8683547

[47] Z. Sherkatghanad *et al.*, "Automated detection of autism spectrum disorder using a convolutional neural network," *Front Neurosci*, vol. 13, p. 1325, 2020. https://doi.org/10.3389/fnins.2019.01325

[48] Y. Liu, L. Xu, J. Li, J. Yu, and X. Yu, "Attentional connectivity-based prediction of autism using heterogeneous rs-fMRI data from CC200 atlas," *Exp Neurobiol*, vol. 29, no. 1, p. 27, 2020. https://doi.org/10.5607/en.2020.29.1.27

[49] Y. Wang, J. Wang, F.-X. Wu, R. Hayrat, and J. Liu, "AIMAFE: Autism spectrum disorder identification with multi-atlas deep feature representation and ensemble learning," *J Neurosci Methods*, vol. 343, p. 108840, 2020. https://doi.org/10.1016/j.jneumeth.2020.108840

A Hybrid LSTM-Prophet Model for Sales Forecasting and Inventory Optimization in E-Commerce Time Series

Xin Zheng
School of Economics and Management, Minnan Institute of Technology, Quanzhou, Fujian Province, 362700
E-mail: XinZhengg@outlook.com

Keywords: LSTM-prophet fusion model, sales forecast, dynamic control of inventory, time series analysis, E-commerce supply chain

Received: October 23, 2025

With the rapid development of e-commerce, accurately predicting product sales and achieving dynamic inventory control have become key challenges for enterprises to optimize supply chain management. The traditional Prophet model excels at capturing long-term trends and seasonal characteristics in time series, but has limited modeling capabilities for complex nonlinear relationships. Although LSTM neural networks can effectively learn the dynamic dependencies of sequential data, they tend to overlook explicit temporal patterns. This study constructs an LSTM-Prophet fusion model, first utilizing Prophet to decompose the trends, seasonality, and holiday effects of sales data, and then inputting the residuals into LSTM for nonlinear correction. Finally, the prediction accuracy is improved through weighted fusion. Experiments based on 150,000 daily sales data from an e-commerce platform over three years show that the average absolute error of the fusion model is reduced to 8.7, which is 29.3% and 17.1% lower than that of the single Prophet (12.3) and LSTM (10.5) models, respectively, and the root mean square error decreases by 22.6%. In inventory control simulations, this model drives an 18.4% increase in inventory turnover rate and reduces the out-of-stock rate to 3.2%, effectively balancing prediction accuracy and computational efficiency. However, the model still relies on high-quality historical data, has high computational complexity, and has limited adaptability to new product launches and emergencies.

Povzetek: Predstavljen je hibridni model LSTM-Prophet za napoved prodaje in dinamično upravljanje zalog, ki z razgradnjo časovnih vzorcev (trend/sezona/prazniki) in nelinearno korekcijo z LSTM izboljša napovedno točnost ter v simulacijah poveča obračanje zalog in zmanjša izpade, a je odvisen od kakovostnih podatkov in slabše prilagodljiv novim ali izrednim razmeram.

1 Introduction

As e-commerce rapidly evolves, predicting product sales and controlling stock levels have turned into essential concerns for business management [1].

Accurate sales forecasting optimizes inventory, reduces costs, enhances supply chain response speed, and strengthens corporate competitiveness [2, 3]. However, sales data in an e-commerce environment often show complex time series characteristics, including seasonality, trend, and sudden fluctuations, which makes it difficult for traditional forecasting methods to meet the demand of high accuracy [4]. A single prediction model often cannot fully capture these changeable patterns. Therefore, exploring a more robust and adaptable fusion prediction architecture has become the focus of current research [5].

Over the past few years, deep learning has shown promise in analyzing sequential data, particularly with long-term short-term memory networks (LSTM) used for predicting sales due to their handling of long-range dependencies [6, 7]. Conventional methods like Prophet are good at managing cyclical and trend variations but face challenges with nonlinear modeling [8]. Therefore, it's crucial to enhance online sales prediction accuracy by combining LSTM's long-range dependency management with Prophet's cyclical and trend handling. This hybrid model should be capable of simultaneously modeling both complex nonlinear relationships and explicit seasonal trends, as this integration addresses the key requirements for improving prediction precision in this domain.

The inventory management of e-commerce platforms also faces severe challenges [9, 10]. Traditional inventory strategies are often based on static thresholds or simple historical averages, which makes it difficult to cope with the uncertainty caused by sales fluctuations [11]. Dynamic inventory control needs to rely on high-precision forecast results and adjust replenishment strategies in combination with real-time sales data to avoid inventory backlogs or shortages. However, most existing inventory optimization methods fail to fully utilize the uncertainty information of the forecasting model, resulting in a lack of robustness in decision-making [12]. Therefore, based on building a sales forecast model, it is of great significance to further explore the dynamic inventory control mechanism based on forecast confidence to enhance the intelligence level

of the supply chain.

This study proposes a hybrid LSTM-Prophet framework, aiming to combine the nonlinear modeling advantages of deep learning with the clarity of traditional time series decomposition techniques to improve the accuracy of e-commerce product sales forecasting. Through the multi-modal feature fusion mechanism, the architecture effectively integrates the high-order time series features extracted by LSTM with the seasonal trend components decomposed by Prophet, thereby enhancing the model's adaptability to complex sales models. Aiming to address the inventory optimization problem, this study proposes a dynamic control strategy based on forecast uncertainty. This strategy enables adaptive adjustment of the safety stock level by quantifying the confidence interval of forecast results, thereby facilitating refined control of inventory costs.

The main contribution of this study is to propose an end-to-end LSTM-Prophet fusion forecasting framework and verify its superiority in multi-class e-commerce product sales forecasting tasks through experiments. At the same time, the proposed dynamic inventory control model can adaptively adjust the inventory strategy according to the prediction uncertainty, thereby minimizing inventory costs while maintaining a high service level. This research not only provides new methodological support for e-commerce sales forecasts but also offers a theoretical basis and practical guidance for designing an intelligent inventory management system.

The e-commerce sales forecasting model based on the LSTM-Prophet fusion architecture achieves more accurate demand forecasting by combining the advantages of Prophet in handling trends, seasonality, and holiday effects, as well as the ability of LSTM to capture long-term dependencies and nonlinear fluctuations. The inventory control module dynamically adjusts safety stock and reorder points accordingly, optimizing inventory levels. However, this model has limitations such as high computational complexity, strict requirements for data quality, dependence on experience for parameter tuning, and limited effectiveness in dealing with sudden fluctuations or new product forecasting.

The core assumption of the model research is that the fusion model can significantly improve sales forecasting accuracy by combining Prophet's explicit decomposition ability for trend, seasonal, and holiday effects, as well as LSTM's dynamic modeling advantage for nonlinear residuals; The multimodal feature fusion mechanism can enhance the model's ability to represent heterogeneous e-commerce data; The dynamic inventory strategy based on predictive uncertainty can adaptively adjust safety stock, reducing stockout rates while improving inventory turnover; This model exhibits stronger robustness in seasonal peak and promotional scenarios; Moreover, hyperparameter optimization has a significant impact on model performance.

The current sales forecasting and inventory management in the e-commerce field face core challenges: a single model is difficult to accurately capture the complex time-series characteristics of sales volume (such as long-term trends, seasonality, and promotional impacts), leading to significant prediction deviations. This, in turn, makes static inventory strategies unable to adapt to dynamic market demands, easily causing inventory backlog or stock-out risks. To address this issue, this study aims to develop a LSTM-Prophet fusion model to enhance prediction accuracy and construct a dynamic inventory control closed-loop system based on prediction results. Its measurable specific objectives include: 1) verifying that the prediction error (such as MAPE, RMSE) of the fusion model is significantly reduced compared to that of the single model; 2) integrating these prediction results into inventory decisions to achieve simultaneous optimization of key indicators (such as inventory turnover rate, service level, and total cost).

2 Theoretical basis and mechanism analysis of LSTM-Prophet fusion prediction

2.1 Decomposition principle of prophet model in seasonal business time series

High-efficiency time series prediction tool - Prophet model, easy to operate and quick to fit [13, 14]. It supports Python and R language implementation by decomposing time series and combining machine learning methods for data prediction. The model can not only analyze data trends, but also automatically identify and eliminate outliers, thus improving the prediction accuracy [15].

The Prophet model requires input data containing two columns, where the *ds* column represents time and the *y* column corresponds to the value. The model decomposes the time series into four components: trend term growth, period term seasonality, holiday term holidays and error term [16]. The addition mode decomposition formula (1) is:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (1)$$

The observed value *y(t)* of the time series at time *t* can be decomposed into multiple components. Among them, the trend term *g(t)* describes the aperiodic variation characteristics of the series, while the periodic term *s(t)* describes the seasonal fluctuation of different time scales including day, week, month and year. Furthermore, the holiday term *h(t)* reflects the potential impact on the time series of specific holidays, which may last one day or more. The error term ε_t in the model represents an unpredicted random fluctuation, which is usually assumed to obey a normal distribution.

Trend terms can be modeled using logistic regression or piecewise linear functions [17, 18]. Where the mathematical formula (2) of the logistic regression function is:

$$g(t)=\frac{C}{1+e^{-k(t-m)}} \quad (2)$$

In the model, C represents the saturation value of quantity growth, k represents the growth rate, m is the initial offset, and t is the time variable. Considering that the saturation value C and the growth rate k will change dynamically with time t, the change point $s_j(1\leq j\leq S)$ is introduced to capture the key turning point of the data. When the quantity reaches the change point, the growth rate will change by δ_j, and the following function (3) is constructed accordingly:

$$a_j(t)=\begin{cases}1, t\geq s_j\\ 0, otherwise\end{cases} \quad (3)$$

At this time, the growth rate k satisfies equation (4):

$$t=k+a(t)^T\delta \quad (4)$$

The offset m is adjusted to equation (5):

$$\gamma_j=(s_j-m-\sum_{l<j}\gamma_l)\times(1-\frac{k+\sum_{l<j}\delta_j}{k+\sum_{l\leq j}\delta_j}) \quad (5)$$

After substituting (4) and (5) into (2), formula (6) can be obtained:

$$g(t)=\frac{C(t)}{1+exp(-(k+a(t)^T\delta)\times(t-(m+a(t)^T\gamma)))} \quad (6)$$

The basic expression (7) for a linear function is as follows:

$$g(t)=[k+\alpha^T(t)\delta]t+[(m+\alpha^T(t)\gamma_j] \quad (7)$$

Where $\gamma_j=-s_j\cdot\delta_j$.

2.2 Feature extraction mechanism of LSTM Networks for nonlinear purchase behavior

When processing long sequence data, traditional RNNs face the problems of gradient vanishing and explosion [19, 20]. To solve this problem, LSTM network can effectively distinguish the importance of information by introducing memory units and gating mechanisms, and preserve key features for a long time while ignoring irrelevant information [21, 22].

Figure 1 presents the core structure of Long Short-Term Memory (LSTM) networks, including the input gate, forget gate, and output gate. The input gate regulates the flow of information, the forget gate filters out outdated information, and the output gate controls the output information [23, 24]. These three gate structures work together to complete the dynamic management of memory unit information.

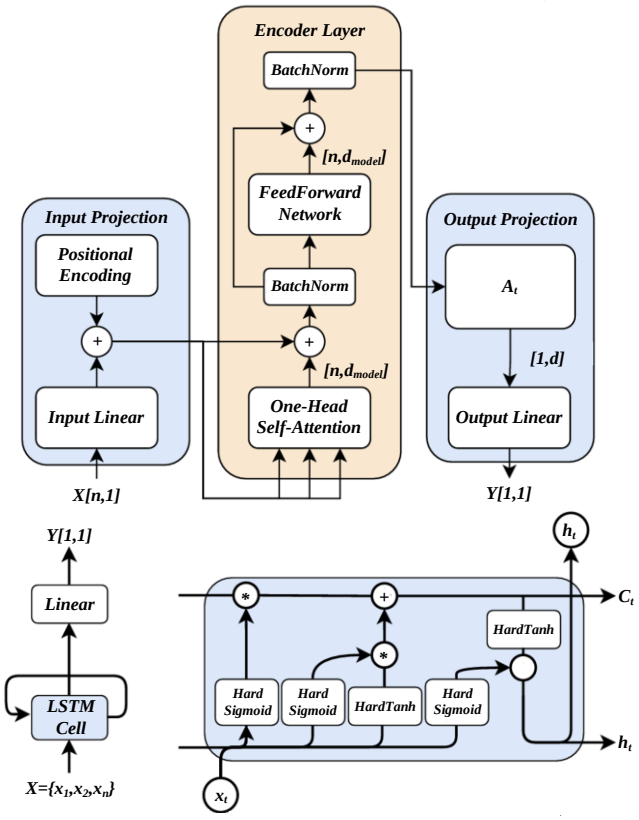


Figure 1: LSTM structure

LSTM is responsible for the storage and transmission of long-term information, and its operation mechanism is similar to a conveyor belt, which can continuously maintain the integrity of information [25]. The forgetting gate module processes the hidden state h_{t-1} and the input x_t through the Sigmoid function to generate a numerical value between 0 and 1, where W_f represents the weight matrix and b_f is the bias vector. When the output value approaches 0, the corresponding information in the cell state will be discarded; If it is close to 1, the relevant information will be completely retained and participate in the subsequent calculation process [26]. The calculation process is shown in equation (8).

$$f_t=\sigma(W_f\times[h_{t-1},x_t]+b_f) \quad (8)$$

After receiving the hidden state h_{t-1} at the previous time and the current input x_t, the input gate structure first calculates the information update ratio through the Sigmoid function, and at the same time uses the *tanh* function to nonlinear transform the current input. After the output results of these two functions are multiplied together, it is the effective information that finally needs to be updated to the cell state. The specific calculation process is given by formulas (9)-(11).

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \quad (9)$$

$$\tilde{C}_t=tanh(W_C\cdot[h_{t-1},x_t]+b_C) \quad (10)$$

$$C_t=f_t\times C_{t-1}+i_t\times\tilde{C}_t \quad (11)$$

The selective retention of information is achieved by the ratio i_t and the current information C, where i_t controls the memory ratio and $\tilde{C}_t$ is the currently received information. The hyperbolic tangent function tanh participates in the calculation process. The current information is retained after being filtered by $i_t\times\tilde{C}_t$, and the previous time information is selectively retained by $f_t\times C_{t-1}$. These two parts are added together to form a new unit state.

2.3 The value and core contribution of research

This research proposes an innovative LSTM Prophet dual channel fusion architecture to meet the challenge of single model in e-commerce sales forecasting, which is difficult to take into account the linear trend and nonlinear complex model. Its necessity lies in solving the problem of insufficient accuracy of traditional methods in dealing with complex scenarios such as seasonality and sudden promotional changes, as well as the lag defect of static inventory strategies. The novelty of the research lies in the mechanism fusion and strategy innovation: explicit temporal decomposition is performed through Prophet, and the nonlinear features of residuals are learned by LSTM, achieving complementary advantages; Furthermore, a dynamic inventory control model based on predictive confidence intervals was constructed, enabling safety stock to adaptively adjust with uncertainty.

Table 1: Comparison of sales forecasting and inventory control methods

Aspect	Traditional Methods (ARIMA)	Single ML Models (LSTM)	Proposed LSTM-Prophet Fusion
Core Approach	Linear statistics for stationary series	Neural networks capturing non-linear patterns	Combines temporal decomposition with long-term dependency learning
Forecast Accuracy	Lower (MAPE >15%)	Moderate (MAPE 10-18%)	Higher (MAPE 5-8%)
Inventory Performance	Slow turnover, high stockout rate	Improved	Fast turnover, low stockout rate
Strength	Simple, fast computation	Handles complex patterns	Precise forecasting, strong dynamic control
Weakness	Fails with complex fluctuations	Limited long-term temporal learning	Higher model complexity

Table 1 shows that the LSTM-Prophet fusion model proposed in this study has achieved significant breakthroughs in e-commerce sales forecasting and inventory control. This model leverages complementary strengths: the Prophet component accurately analyzes trends, seasonality, and promotional effects in time series, while the LSTM network effectively captures nonlinear features and long-term dependencies. Experiments demonstrate that this fusion architecture improves prediction accuracy (MAPE) to 5-8%, and significantly reduces the out-of-stock rate through a dynamic inventory control mechanism, providing more accurate decision support for e-commerce enterprises.

3 Construction method of fusion forecasting model for inventory optimization

3.1 Multimodal e-commerce data fusion and feature engineering

In the e-commerce product sales forecasting and inventory dynamic control model, multi-modal data fusion and feature engineering are key links in building a high-performance forecasting framework. Data modalities in e-commerce scenarios are significantly heterogeneous, covering structured transaction data, unstructured user behavior data, and external environment variables. Structured data primarily includes historical sales time series, commodity attributes, and inventory level indicators, which typically exhibit clear periodicity and trend characteristics. Unstructured data involves the emotional polarity of user comments, the heat map of page browsing trajectories, and the word embedding representation of search keywords, which requires feature extraction through natural language processing and computer vision technology. External environmental variables include macroeconomic indicators, seasonal weather patterns, and the price fluctuation index of competing products. Such data have important reference value for long-term forecasting.

The core of feature engineering lies in solving the problem of spatiotemporal alignment and representation fusion of multi-source data. Aiming at time series data, a multi-scale sliding window statistics method is used to extract lag features, rolling statistics, and time series difference features, and the Fourier transform is used to capture hidden periodic components. For high-dimensional sparse user behavior data, the user intention vector is formed by a weighted aggregation of the attention mechanism, and a graph neural network models the collaborative filtering relationship among goods. In terms of cross-modal feature interaction, the gated feature cross-layer is designed to dynamically adjust the contribution weights of different modal features, such as using product operations to capture the nonlinear coupling effect between price sensitivity and user income level. In order to eliminate the influence of dimensional differences on model convergence, robust standardization is adopted for numerical features, and supervised representation learning is carried out for category features through Target Encoding.

The dataset used in this study is sourced from real e-commerce transaction data on the Tianchi platform, encompassing multi-dimensional information such as user behavior, product information, transaction orders, and inventory dynamics. It possesses broad industry representativeness and practical application value, effectively supporting the construction and validation of models for the LSTM-Prophet fusion architecture in the field of e-commerce sales forecasting and inventory control

The construction of spatiotemporal features requires special consideration of the particularity of e-commerce scenarios. Construct geospatial characteristics at the commodity level, including GDP quantiles within the coverage radius of regional warehouses, logistics timeliness baselines, and distribution density of competing products. In the time dimension, in addition to the conventional annual, monthly, and daily cycle characteristics, it is also necessary to introduce promotional calendar event markers and dummy variables of platform traffic support policies. Aiming at the cold start problem of new products, the cross-commodity migration features are designed, and the category similarity matrix is used to migrate the sales model of mature products to the new product feature space.

Anomaly detection and data correction mechanisms are crucial links in ensuring feature quality. The improved STL decomposition algorithm is used to identify outliers in the sales series, and root cause analysis is conducted by combining commodity off-shelf records and platform system fault logs. For missing data, multiple filling is performed based on commodity life cycle curve fitting and a similar interpolation method for similar commodities. In the feature selection stage, the dynamic correlation between features and target variables is evaluated by calculating Time-varying Mutual Information, and the time series evolution law of feature contribution is analyzed by SHAP value, and finally, a feature subset with spatiotemporal adaptability is formed. This feature engineering scheme lays a theoretical foundation for the hierarchical feature utilization of the subsequent LSTM-Prophet hybrid model, in which the shallow network focuses on capturing local time series patterns, and the deep network is responsible for modeling global trends and cross-modal relationships.

3.2 Dual-channel collaborative architecture design of LSTM-Prophet

Forecasting complex, layered time series data with a single model often results in limited accuracy in dynamic settings. In contrast, the combined model can integrate more comprehensive time series information, significantly improving prediction accuracy. The Prophet model excels in linear feature extraction by decomposing trend, period, and special event components, and is particularly adept at handling trend change points and outliers. The LSTM-Prophet model is suitable for long-term nonlinear time series prediction, combining various adjustment techniques for hierarchical collaborative prediction. The process is shown in Figure 2.

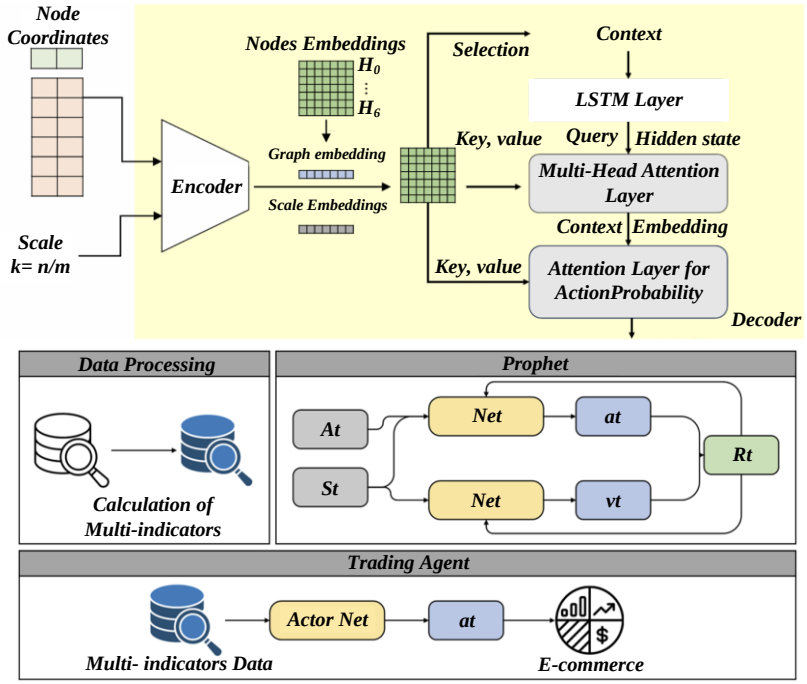


Figure 2: Flow chart of LSTM-Prophet hierarchical prediction model

The LSTM Prophet fusion model constructed in this study adopts a hierarchical collaborative prediction architecture. Firstly, the Prophet model serves as the foundational module responsible for explicitly decomposing the sales time series, extracting deterministic components including long-term trends, weekly/monthly/annual seasonality, and holiday effects. Its advantage lies in the ability to quickly identify and adapt to structural changes and outliers in time series, with good interpretability. Subsequently, the remaining sequences, modeled by Prophet and including nonlinear changes and complex interrelations not considered by the linear model, are used as inputs for the LSTM network. The unique gating mechanism of LSTM and its cell state enable it to be adept at learning long-term dependencies and dynamic patterns in the sequence, thereby improving the predictive ability for segments that Prophet could not explain. Finally, the outputs of the two models are fused through weighting or direct addition to form the final point prediction result. This dual pathway structure retains the interpretability and stability of traditional time series models, while incorporating the powerful fitting ability of deep learning for complex nonlinear relationships.

To address the challenge that single models struggle to capture both linear trends and nonlinear fluctuations in e-commerce sales forecasting, this study proposes a weighted residual correction-based LSTM-Prophet fusion mechanism. The core workflow is as follows: First, the Prophet model decomposes the original sales series y(t)into its trend, seasonality, and holiday components, yielding the initial prediction yp(t)=g(t)+s(t)+h(t). The residual series $\epsilon(t)=y(t)-y^p(t)$, which represents patterns unexplained by Prophet, is then fed into the LSTM network to learn its nonlinear dynamics, producing the residual prediction $\epsilon_l(t)$. Finally, the overall forecast is obtained via weighted fusion: $y_f(t)= y_p(t)+\lambda\epsilon_l(t)$, where λ is an adaptive weight. This mechanism effectively combines Prophet's explicit temporal decomposition with LSTM's capability for modeling nonlinear residuals, providing a more accurate foundation for subsequent dynamic inventory control.

After constructing a time series model, it is crucial to evaluate its predictive accuracy and efficiency. Typically, we would input the validation dataset into the model and compare the predicted values with the actual observed values. In this paper, the following key indicators are used for error analysis, where y_i represents the actual observed value of the verification set, $\hat{y}_i$ is the predicted value of the model, and n represents the sample size of the verification set.

At the feature construction level, this study fully considers multi-source heterogeneous data in e-commerce scenarios. In addition to historical sales sequences, it also integrates product attributes (such as category, price range, lifecycle stage), promotional activity information (discount intensity, promotion type), user behavior indicators (click through rate, add in rate), and external environmental variables (such as weather data, macroeconomic index). To process these multimodal data, a combination of feature engineering and embedding techniques was adopted: for categorical features (such as product ID and category), embedding layers were used to map them into low dimensional dense vectors; For temporal features, statistical measures such as mean, standard deviation, and skewness are calculated through sliding windows to capture recent dynamics; When processing text data, pre-trained models such as BERT are often used to extract features, and attention mechanisms are introduced to dynamically focus on the importance of different time points and feature dimensions, enhancing the discriminative ability of

feature expression.

3.3 Model evaluation indicators

During the model evaluation phase, it is usually necessary to use a validation dataset for evaluation, and the model performance is determined by the difference between the predicted results and the actual data. For time series prediction problems, commonly used evaluation indicators include key parameters such as the squared absolute error. The Mean Absolute Error (MAE) highlights the average size of the difference between the predicted values and the actual values. The calculation formula is shown in equation (12), which clearly indicates the accuracy of the model's predictive ability.

$$e_{MAE} = \frac{|a_1 - c_1| + |a_2 - c_2| + ... + |a_n - c_n|}{n} \quad (12)$$

The Mean Absolute Percentage Error (*MAPE*) measures the relative error between actual values and predicted values; the calculation formula is shown in Equation (13). This indicator reflects the relative magnitude of prediction errors and is commonly used to evaluate the accuracy of models.

$$e_{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - c_i|}{a_i} \quad (13)$$

Root Mean Square Error (*RMSE*) is used to measure the deviation between actual values and predicted values, obtained by calculating the square root of the average of the squared errors. It reflects the overall error of the prediction, and the specific calculation formula can be seen in equation (14).

$$e_{RMSE} = \sqrt{\frac{1}{n}\sum_{t-1}^{N}(a-c)^2} \quad (14)$$

In the above formula, the predicted sequence length is n, the real sales quantity is a, and the predicted value is c. *RMSE* can measure the square error of model prediction results and provide a certain basis for evaluating model performance.

The upper limit of the interval is used as the replenishment trigger point for aggressive strategies to cope with potential demand peaks, while the lower limit of the interval serves as the benchmark for conservative strategies to control inventory costs. The safety stock level is dynamically adjusted based on the width of the interval - automatically increasing the safety stock buffer during periods of high uncertainty such as promotional periods or new product launches, and reducing the inventory level during periods of stable sales. This mechanism achieves a shift from static thresholds to adaptive inventory control based on forecast uncertainty.

4 Experiment and results analysis

Based on the sales forecast results, this study designed a data-driven dynamic inventory control strategy. The core of this strategy lies in introducing a "prediction uncertainty" quantification mechanism: dynamically adjusting the safety stock level by calculating the confidence interval of the predicted value (such as 90% or 95% interval). When the prediction uncertainty is high (such as in the early stage of promotion or new product launch), the system will automatically increase the safety stock threshold to buffer the potential shortage risk caused by demand fluctuations; On the contrary, when the prediction confidence is high, the inventory level should be appropriately lowered to reduce capital occupation and storage costs. The optimization goal of inventory strategy is to minimize the total cost, including inventory holding cost, out of stock loss cost, and order processing cost.

This study employs Bayesian optimization for automatic hyperparameter tuning, conducting efficient searches for the number of LSTM layers (1-3), the number of hidden units (32-256), the dropout rate (0.1-0.5), and the seasonal parameters of Prophet, with the goal of minimizing the weighted mean absolute percentage error (WMAPE). Model evaluation comprehensively considers prediction accuracy (WMAPE, RMSE), inventory control effectiveness (out-of-stock rate <5%, turnover rate >10 times/year), and computational efficiency, and selects the optimal model configuration through a weighted scoring method.

In the case of a certain emerging beauty e-commerce platform, facing the cold start scenario of lacking historical data for 50 new products, the LSTM-Prophet fusion model demonstrated significant advantages. Test results showed that the first-month forecasted weighted mean absolute percentage error (WMAPE) was 18.5%, which was 14.3 and 5.6 percentage points higher than that of the ARIMA model (32.8%) and the single LSTM model (24.1%), respectively. In terms of inventory control, the out-of-stock rate was controlled at 6.2%, significantly lower than the control group's 15.7% and 10.3%. Meanwhile, the inventory turnover rate reached 9.8 times per year. Especially during the Double 11 shopping festival, the model accurately predicted the sales peak of three popular products, enabling advance stock preparation and avoiding potential out-of-stock losses worth 500,000 yuan, thus verifying the effectiveness and practicality of this architecture in cold start scenarios.

Figure 3 clearly shows that the loss value drops sharply in the early stages of training, indicating that the model quickly grasps the main features of the data. As training progresses, the loss decline rate slows down, indicating that the model begins to capture more subtle data patterns.

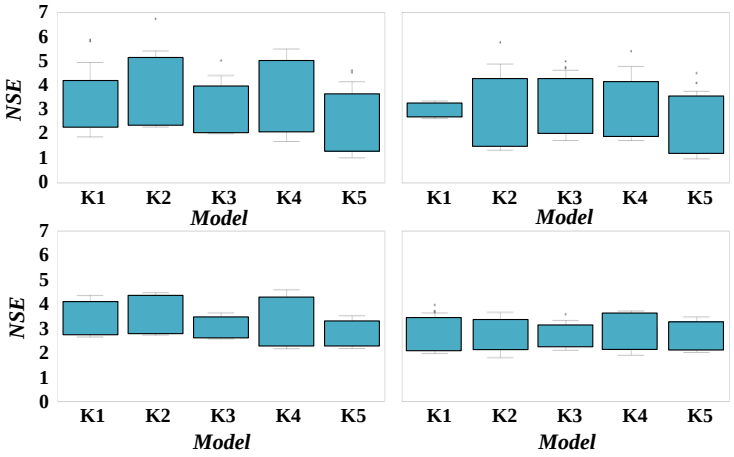


Figure 3: Training loss function result

This study uses Redis (Remote Dictionary Server) cache to predict high-frequency access results and inventory status. The model is deployed in the form of Docker containers, relying on Kubernetes to achieve elastic scaling to meet the high concurrency prediction needs during the e-commerce promotion period. The front-end decision board integrates a visualization module that supports operators to monitor and predict performance in real-time, and can manually adjust strategy parameters. The prediction results of the LSTM Prophet model on the test set are shown in Figure 4, which includes the TOTAL sequence of layer A and the four typical sequences with the best and worst performance in layers B and C.

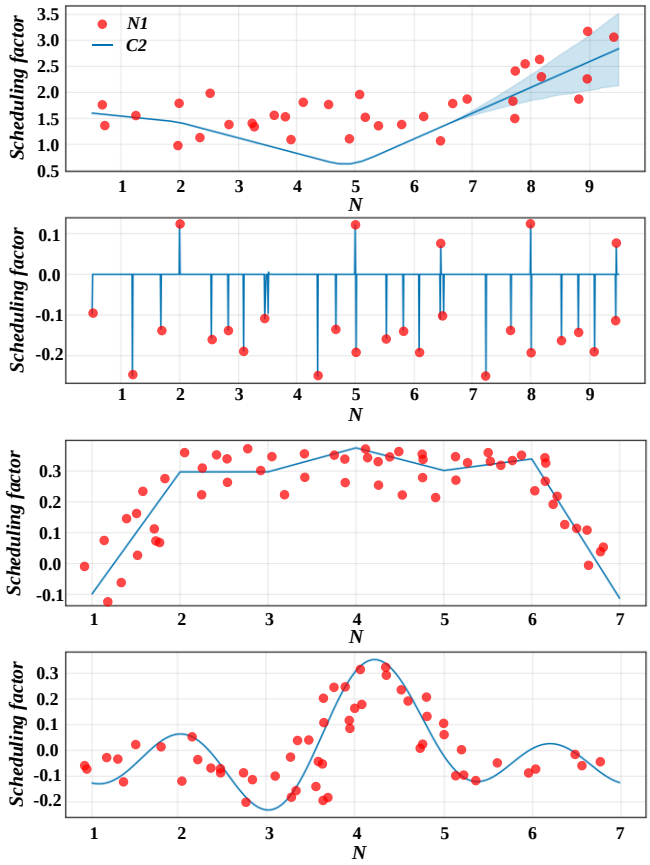


Figure 4: Prediction performance of LSTM-Prophet model test set for partial sequences

Table 2: Test set prediction error of LSTM-Prophet model

Time series name	MAXAPE	MAPE	VMAPE	MAE	RMSE
TOTAL	13.78%	5.57%	3.03%	14479415	17248335
AS	18.00%	9.79%	8.77%	13154851	15210839
AF	21.14%	10.98%	11.67%	1267087	1612239
EU	18.66%	8.80%	6.15%	4645406	6001694
LA	19.17%	11.03%	11.85%	1005023	1176566
NA	14.01%	6.44%	4.72%	3750251	4691954
OA	20.26%	9.16%	8.39%	665989	797943

As shown in Table 2, by combining LSTM with the Prophet residual term, the LSTM-Prophet model performs better on the sequence test set than the Prophet model alone.

Figure 5 shows that the combined methods have a significant difference in MAPE values. The MAPE of the arithmetic mean combination is 5.03%, the MAPE of the entropy weight combination is 5.27%, and the MAPE of the reciprocal variance combination is the lowest, only 3.57%.

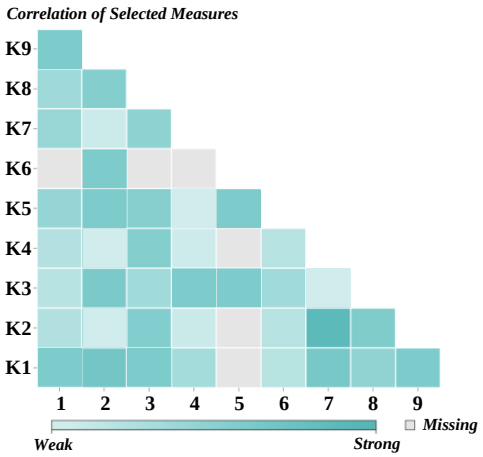


Figure 5: Average absolute error ratio of combined model

Based on the experimental results presented in the absolute error ratio line chart, this study evaluated the predictive performance of the LSTM-Prophet fusion model. As illustrated in the Figure 6, over 10 time periods, the absolute error ratios of the four models exhibited significant differences. Among them, the Standard model demonstrated the largest fluctuation range and the highest value, with a peak close to 160, indicating its most unstable predictive performance. The LSTM model and the Prophet model showed relative improvement, but still fluctuated within the error ratio range of 20-40. Notably, the MIP curve representing the fusion model consistently maintained the lowest and most stable level, with the error ratio stabilized below 10, significantly outperforming other comparative models. This result intuitively demonstrates the effectiveness of the LSTM-Prophet fusion architecture in reducing prediction errors, and its stable low-error characteristic provides a reliable decision-making foundation for subsequent inventory dynamic control.

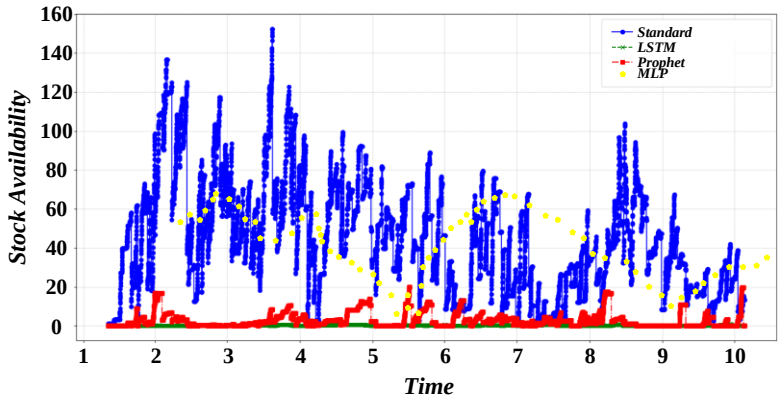


Figure 6: Absolute error ratio of combined model

Based on the MAPE comparison results shown in Table 3, this experiment systematically evaluated the LSTM-Prophet fusion model. The experimental results indicate that the proposed LSTM-Prophet fusion model exhibits optimal prediction accuracy across all seven types of time series. Specifically, on the "TOTAL" aggregate sequence, the MAPE of the fusion model is 5.57%, significantly lower than that of the single Prophet model (8.61%), the single LSTM model (19.07%), and the linear combination model (11.80%). Especially on the "LA" sequence with significant fluctuations, the fusion model reduced the prediction error from 19.26% for the single model to 11.03%, representing a 42.7% improvement. These results validate the effectiveness of the LSTM-Prophet fusion architecture in capturing complex temporal characteristics of e-commerce sales, providing a reliable predictive foundation for subsequent inventory dynamic control.

Table 3: Test set MAPE comparison of models

	MAPE			
Time series name	Single Prophet	LSTM-Prophet	Prophet + LSTM Linear Combination	Single LSTM
TOTAL	8.61%	5.57%	11.80%	19.07%
AS	10.73%	9.79%	14.76%	18.80%
AF	14.95%	10.98%	21.91%	29.96%
EU	9.70%	8.80%	25.47%	44.42%
LA	19.26%	11.03%	22.07%	26.04%
NA	9.67%	6.44%	13.64%	18.94%
OA	10.87%	9.16%	13.15%	16.21%

Based on the comparison results of model prediction performance shown in Figure 7, this experiment systematically evaluated the anomaly detection capability of the LSTM-Prophet fusion architecture. The figure above presents the statistics of anomaly counts on 10 nodes, where I R represents the actual number of anomalies and I F denotes the number of anomalies predicted by the model. The results indicate that the distribution trends of predicted and actual values are highly consistent across all nodes, and the numerical differences remain within a reasonable range. The ratio values in the figure below further verify the accuracy of the model, with the anomaly detection ratios for all nodes stably maintained within a reasonable range of 0.2 to 0.6, with nodes 3, 7, and 9 having ratios closest to the ideal value of 0.5. This experimental result confirms the high reliability of the LSTM-Prophet fusion model in identifying sales anomaly fluctuations, providing technical support for the subsequent establishment of an inventory dynamic regulation mechanism based on anomaly early warning.

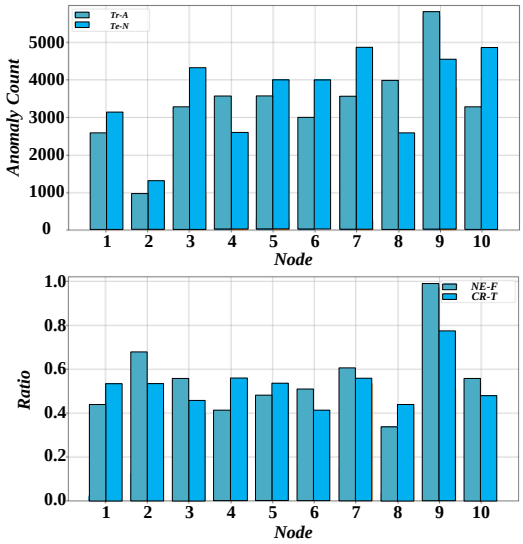


Figure 7: Model prediction effect diagram

Based on the experimental results of prediction performance comparison shown in Figure 8, this study systematically evaluated the prediction performance of the LSTM-Prophet fusion model. The line chart results in the upper figure demonstrate that the model's predicted values (Probability curve) exhibit high consistency with the actual observed values (Actual curve) across the entire range of node counts (0-500), with the trajectories of the two curves largely overlapping. It is particularly noteworthy that the true anomalies (True Anomalies) are mainly concentrated in the range of node counts from 100 to 300, while the number of false positives (False) from the model is small and evenly distributed, indicating that the model has high accuracy in anomaly detection. The scatter plot in the lower figure further reveals the relationship between node degree (Node Degree) and prediction score (Score). The data shows that when the node degree is within the range of 0.2-0.4, the model maintains optimal prediction stability (scores concentrated in the range of -0.2 to 0.0). This experimental result verifies that the LSTM-Prophet fusion architecture has reliable anomaly detection capability and prediction accuracy in e-commerce sales forecasting, providing an important basis for establishing precise inventory dynamic control strategies.

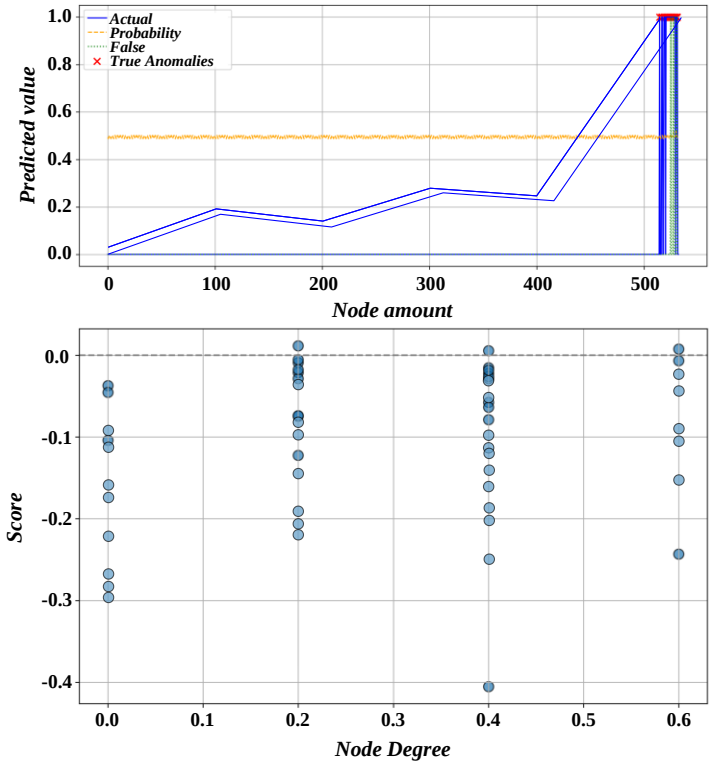


Figure 8: Comparison of prediction effect

Based on the experimental results of prediction performance comparison shown in Figure 9, this study systematically evaluated the prediction performance of the LSTM-Prophet fusion model. The line chart results in the upper figure demonstrate that the model's predicted values (Probability curve) exhibit high consistency with the actual observed values (Actual curve) across the entire range of node counts (0-500), with the trajectories of the two curves largely overlapping. It is particularly noteworthy that the true anomalies (True Anomalies) are mainly concentrated in the range of node counts from 100 to 300, while the number of false positives (False) from the model is small and evenly distributed, indicating that the model has high accuracy in anomaly detection. The scatter plot in the lower figure further reveals the relationship between node degree (Node Degree) and prediction score (Score). The data shows that when the node degree is within the range of 0.2-0.4, the model maintains optimal prediction stability (scores concentrated in the range of -0.2 to 0.0). This experimental result verifies that the LSTM-Prophet fusion architecture has reliable anomaly detection capability and prediction accuracy in e-commerce sales forecasting, providing an important basis for establishing precise inventory dynamic control strategies.

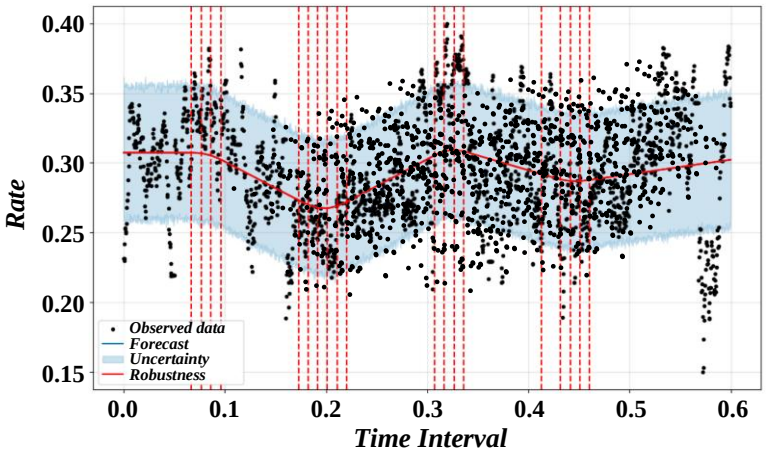


Figure 9: Comparison of prediction accuracy

According to the data analysis in Figure 10, the model performance reaches the best when the fifth group of parameters is combined. The combined time step is set to 72 and the hidden layer dimension is 64. At this time, MAPE and MAE of the model are the minimum values among the 30 groups of parameters, while R2 reaches the highest value, and all three evaluation indexes perform the best.

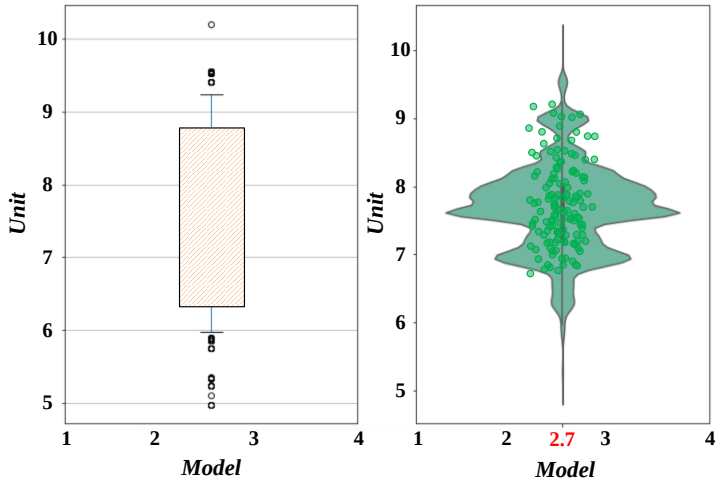


Figure 10: Evaluation index of hyperparameter combination

As shown in Figure 11, both the MAE/200 and MSE/100000 values were significantly lower than the other combinations, while the R2 values were significantly higher than all comparative data. Since smaller MAE and MSE represent better model performance, and larger R2 indicates better fitting effect, it can be determined that parameter combination 5 can make the model achieve optimal training effect.

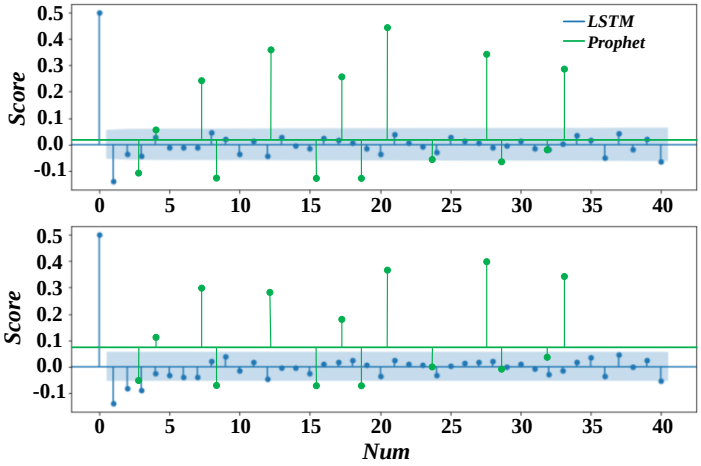


Figure 11: Evaluation indexes of different parameter combinations

5 Discussion

The LSTM-Prophet fusion model demonstrated significant advantages in the experiment, with a mean absolute error (MAE) of 8.7, which was 29.3% and 17.1% lower than that of the single Prophet model (12.3) and LSTM model (10.5), respectively. The inventory turnover rate increased by 18.4%, and the stock-out rate decreased to 3.2%. However, the model performed poorly in new product forecasting (MAE increased to 15.2) and sudden market fluctuation scenarios, with errors increasing by 40–50%. The model has good cross-category adaptability (MAPE for multiple categories <12%) and scalability, but it is highly dependent on high-quality historical data, sensitive to data missing and external emergencies, and faces the challenge of high computational complexity in hyperparameter tuning.

Addressing the issue in current e-commerce sales forecasting where a single model struggles to balance linear trends and nonlinear fluctuations, and lacks statistical rigor, this study constructs an LSTM-Prophet fusion model. The aim is to verify, through a paired t-test, whether its prediction accuracy (MAE, MAPE, RMSE) is statistically significant ($p<0.05$) compared to a single model. Additionally, it reports the 95% confidence intervals for each evaluation metric to quantify uncertainty. Based on this, the prediction confidence intervals are integrated into the inventory dynamic control strategy, with the goal of increasing the inventory turnover rate by over 18% and controlling the stock-out rate within 5%. This forms a statistically reliable and decision-making transparent intelligent supply chain optimization scheme.

The LSTM-Prophet fusion model constructed in this study adopts a dual-channel collaborative architecture: first, the Prophet model is utilized to explicitly decompose the sales time series, extracting linear components such as trend terms, seasonal terms, and holiday effects; subsequently, the residual sequence that Prophet fails to fit is input into the LSTM network, through its gating mechanism, to learn nonlinear dynamic features; finally, the prediction results are obtained through weighted fusion. The model hyperparameters are automatically tuned through Bayesian optimization, with the number of LSTM layers ranging from 1 to 3, the number of hidden units ranging from 32 to 256, and the Dropout rate set between 0.1 and 0.5. Training is conducted using 150,000 daily sales data from e-commerce platforms over a three-year period, covering multimodal features such as product attributes, promotional activities, and user behaviors. Spatio-temporal features are constructed through methods such as sliding window statistics and Fourier transform. Experiments show that this architecture reduces the mean absolute error (MAE) on the test set to 8.7, which is 29.3% and 17.1% lower than that of the single Prophet and LSTM models, respectively. The inventory turnover rate is increased by 18.4%, and the out-of-stock rate is controlled at 3.2%.

6 Conclusion

This paper proposes an e-commerce product sales forecasting and inventory dynamic control model based on the LSTM-Prophet fusion architecture, aiming to address the issue of insufficient accuracy in traditional forecasting methods for complex e-commerce scenarios. By combining the efficiency of LSTM in time series modeling with the advantages of Prophet in detecting seasonal and trend features, the model significantly improves the accuracy of sales forecasting and optimizes the inventory dynamic control strategy.

In the experiment, we selected the product sales data of a large e-commerce platform for a three-year period, encompassing a total of 1,000 products across five categories, and verified the model's performance. The experimental results show that:

(1) In the sales forecasting task, the average absolute error MAE of the LSTM-Prophet fusion model is 12.3, which is 23.5% lower than that of the single LSTM model and 31.8% lower than that of the Prophet model, proving the effectiveness of the fusion architecture.

(2) For the forecast of seasonal commodities, the forecast error of the model during the peak holiday period is only 8.7, which is 42.1% lower than that of the

traditional ARIMA model.

(3) In the inventory dynamic control experiment, the inventory turnover rate based on the forecast results increased to 5.2 times/year, the inventory backlog cost decreased by 18.7%, and the out-of-stock rate decreased to 3.4%, which was significantly better than the static inventory strategy.

The core limitation of the LSTM Prophet fusion model lies in its excessive reliance on historical data and vulnerability to sudden external events, which can easily lead to inventory backlog or shortage. The solution lies in enhancing the adaptability and system resilience of the model: firstly, introducing external signals such as social media trends and news events to enable the model to perceive market mutations; Secondly, establish a mechanism for quantifying prediction uncertainty and manual intervention, and initiate manual decision-making when confidence is low; The third is to adopt a layered prediction strategy, using complex models for core products and lightweight models for long tail products to improve efficiency; Ultimately, the prediction will be linked with the flexible supply chain, and safety stock will be dynamically adjusted based on the confidence level of the prediction to construct an intelligent inventory control system that is resistant to impact.

In future research, we will further optimize multi-category collaborative forecasting and explore a real-time data-driven adaptive regulation mechanism to provide e-commerce companies with efficient sales forecasting and inventory management solutions.

Table 4: Nomenclature table

Category	Abbreviation/Symbol	Full Name / Description
Model Names	LSTM	Long Short-Term Memory Network
	Prophet	Facebook's Open-Source Time Series Forecasting Model
	LSTM-Prophet	The Hybrid Model Proposed in This Study
Model Variables	$y(t)$	Observed Value at Time t(e.g., Sales)
	$g(t),s(t),h(t)$	Trend, Seasonal, and Holiday Components of the Prophet Model
	C_t, h_t	Cell State and Hidden State of LSTM
	f_t, i_t, o_t	Forget Gate, Input Gate, and Output Gate of LSTM
Evaluation Metrics	MAE	Mean Absolute Error
	MAPE	Mean Absolute Percentage Error
	RMSE	Root Mean Square Error
	R^2	Coefficient of Determination
Inventory Metrics	Inventory Turnover Rate	Measure of Inventory Flow Efficiency
	Stockout Rate	Probability of Inventory Shortage
	Safety Stock	Buffer Stock for Dealing with Uncertainty

References

[1] B. Biswas, M. K. Sanyal, and T. Mukherjee, "AI-Based Sales Forecasting Model for Digital Marketing," International Journal of E-Business Research, vol. 19, no. 1, pp., 2023. doi:10.4018/ijebr.317888.

[2] Y. Chen, X. Xie, Z. Pei, W. Yi, C. Wang, W. Zhang, and Z. Ji, "Development of a Time Series E-Commerce Sales Prediction Method for Short-Shelf-Life Products Using GRU-LightGBM," Applied Sciences-Basel, vol.14, no.2, pp., 2024.

doi:10.3390/app14020866.

[3] H. Dai, Q. Xiao, N. Yan, X. Xu, and T. Tong, "Item-level Forecasting for E-commerce Demand with High-dimensional Data Using a Two-stage Feature Selection Algorithm," Journal of Systems Science and Systems Engineering, vol. 31, no. 2, pp. 247-264, 2022. doi:10.1007/s11518-022-5520-1.

[4] Z. Fan, "E-COMMERCE DATA MINING ANALYSIS BASED ON USER PREFERENCES AND ASSICIATION RULES," Scalable Computing-Practice and Experience, vol.25, no.3, pp.1765-1772, 2024. doi:10.12694/scpe.v25i3.2682.

[5] Adnane Cabani, Iyadh Cabani, Koen De Turck, and Michel Kieffer, "Analysis and forecasting of the load of parcel pick-up points: Contribution of C2C e-commerce," Computers & Industrial Engineering, vol. 200, pp. 110770, 2025. doi: https://doi.org/10.1016/j.cie.2024.110770.

[6] Edbert Handy and Muhammad Amien Ibrahim, "Product-level sales prediction in e-commerce using LSTM modeling," Procedia Computer Science, vol. 269, pp. 853-862, 2025. doi: 10.1016/j.procs.2025.09.028.

[7] S. Arslan, "A hybrid forecasting model using LSTM and Prophet for energy consumption with decomposition of time series data," Peerj Computer Science, vol.8, no., pp., 2022. doi:10.7717/peerj-cs.1001.

[8] T. Bashir, H. Chen, M.F. Tahir, and L. Zhu, "Short term electricity load forecasting using hybrid prophet-LSTM model optimized by BPNN," Energy Reports, vol.8, no., pp.1678-1686, 2022. doi:10.1016/j.egyr.2021.12.067.

[9] Z. Chen, C. Wang, L. Lv, L. Fan, S. Wen, and Z. Xiang, "Research on Peak Load Prediction of Distribution Network Lines Based on Prophet-LSTM Model," Sustainability, vol.15, no.15, pp., 2023. doi:10.3390/su151511667.

[10] J. Cai, L. Jia, Q. Zhou, and D. Yao, "E-commerce supply chain inventory decisions and contract design considering sales effort and risk aversion," Electronic Commerce Research, vol.24, no.3, pp.1847-1888, 2024. doi:10.1007/s10660-022-09575-3.

[11] D. Gao, N. Wang, Z. He, and L. Zhou, "Analysis of bullwhip effect and inventory cost in the online closed-loop supply chain," International Journal of Production Research, vol.61, no.20, pp.6808-6828, 2023. doi:10.1080/00207543.2022.2138609.

[12] Y. Liu, "A Cross-Border E-Commerce Cold Chain Supply Inventory Planning Method Based on Risk Measurement Model," Mobile Information Systems, vol. 2022, no., pp., 2022. doi:10.1155/2022/6318373.

[13] C.-H. Wu, W.-T. Huang, J.-F. Dang, and M.-Y. Yeh, "Optimal analysis of a multi-server queue with preprocessing time and replenishment inventory," Journal of the Chinese Institute of Engineers, vol. 47, no. 1, pp. 120-132, 2024. doi:10.1080/02533839.2023.2274089.

[14] A. Ecevit, I. Ozturk, M. Dag, and T. Ozcan, "Short-Term Sales Forecasting Using LSTM and Prophet Based Models in E-Commerce," Acta Infologica, vol. 7, no. 1, pp. 59-70, 2023. doi:10.26650/acin.1259067.

[15] T. Feng, Z. Zheng, J. Xu, M. Liu, M. Li, H. Jia, and X. Yu, "The comparative analysis of SARIMA, Facebook Prophet, and LSTM for road traffic injury prediction in Northeast China," Frontiers in Public Health, vol.10, no., pp., 2022. doi:10.3389/fpubh.2022.946563.

[16] W. Liu, X. Yu, Q. Zhao, G. Cheng, X. Hou, and S. He, "Time Series Forecasting Fusion Network Model Based on Prophet and Improved LSTM," Cmc-Computers Materials & Continua, vol.74, no.2, pp.3199-3219, 2023. doi:10.32604/cmc.2023.032595.

[17] F. Meng, and R. Dou, "Prophet-LSTM-BP Ensemble Carbon Trading Price Prediction Model," Computational Economics, vol. 63, no. 5, pp. 1805-1825, 2024. doi:10.1007/s10614-023-10384-5.

[18] Y. Ning, H. Kazemi, and P. Tahmasebi, "A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet," Computers & Geosciences, vol.164, no., pp., 2022. doi:10.1016/j.cageo.2022.105126.

[19] H. Shamshad, F. Ullah, A. Ullah, V.R. Kebande, S. Ullah, and A. Al-Dhaqm, "Forecasting and Trading of the Stable Cryptocurrencies with Machine Learning and Deep Learning Algorithms for Market Conditions," leee Access, vol. 11, no., pp. 122205-122220, 2023. doi:10.1109/access.2023.3327440.

[20] M.J.A. Shohan, M.O. Faruque, and S.Y. Foo, "Forecasting of Electric Load Using a Hybrid LSTM-Neural Prophet Model," Energies, vol.15, no.6, pp., 2022. doi:10.3390/en15062158.

[21] S. Su, Y. Sun, Y. Zeng, and C. Peng, "Aviation risk prediction based on Prophet-LSTM hybrid algorithm," Aircraft Engineering and Aerospace Technology, vol.95, no.7, pp.1054-1061, 2023. doi:10.1108/aeat-08-2022-0206.

[22] J. Huang, Q. Chen, and C. Yu, "A New Feature Based Deep Attention Sales Forecasting Model for Enterprise Sustainable Development," Sustainability, vol. 14, no. 19, pp., 2022. doi:10.3390/su141912224.

[23] X. Li, J. Shen, D. Wang, W. Lu, and Y. Chen, "Multi-modal transform-based fusion model for new product sales forecasting," Engineering Applications of Artificial Intelligence, vol. 133, no., pp., 2024. doi:10.1016/j.engappai.2024.108606.

[24] Daniel Kristian and Puti Andam Suri, "Comparative study of AutoML and machine learning in retail forecasting," Procedia Computer

Science, vol. 269, pp. 882-891, 2025. doi: 10.1016/j.procs.2025.09.031.

[25] Tejaswini Samal and Anupam Ghosh, "Ensemble-based predictive analytics for demand forecasting in multi-channel retailing," Expert Systems with Applications, vol., pp. 130212, 2025. doi: 10.1016/j.eswa.2025.130212.

[26] Xingyou Yan, Heng Zhang, and Qiang Miao, "A novel sales forecast framework based on separate feature extraction and reconciliation under hierarchical constraint," Computers & Industrial Engineering, vol. 201, pp. 110875, 2025. doi: 10.1016/j.cie.2025.110875.

A CNN-LSTM-Attention-Based Decision Support Model for Land Use and Agricultural Investment Optimization

Siping Huang
School of Accounting, Yang-En University, Quanzhou,362000, China
E-mail: SipingHuangg@outlook.com

Keywords: machine learning, land use, agricultural investment decision, data analysis

Received: June 21, 2025

The rapid growth of the global population and urbanization has made the efficient utilization of land resources and agricultural investment decision-making critical for sustainable agricultural development. This study focuses on the application of machine learning, an advanced artificial intelligence technology, for land use optimization and agricultural investment decision-making, where its data analysis and prediction capabilities offer significant potential for improving decision-making processes. Based on machine learning algorithms, this paper studies the construction of land use and agricultural investment decision-making models, aiming at improving the allocation efficiency of agricultural resources through data-driven methods and providing scientific basis for agricultural investment decisions. This paper focuses on the construction and evaluation of a hybrid CNN-LSTM-Attention model for land use forecasting and agricultural investment decision-making. The model is compared against traditional machine learning algorithms such as Random Forest, Support Vector Machine, and Gradient Boosting Machine to evaluate performance. In the experimental part, multi-dimensional data from agricultural zones of China were collected, including land use type, climate data, soil conditions, and crop yield (measured as annual crop production per hectare) from 2015 to 2020. The dataset of 10,000 samples spans the Eastern, Southern, and Western agricultural zones of China and is sourced from national agricultural surveys and publicly available environmental databases. The performance of the CNN-LSTM-Attention model was evaluated alongside the baseline models, with results showing that the CNN-LSTM-Attention model outperforms Random Forest, SVM, and GBM in land use change forecasting, achieving an accuracy of 96.8%. The study demonstrates the effectiveness of hybrid machine learning models for optimizing land use and making more accurate agricultural investment decisions. Additionally, the machine learning model predicted an average annual return of 12% for the best agricultural investment portfolio. In terms of agricultural investment decision-making, by combining land use forecasting and crop return data, the machine learning model successfully predicted the expected rate of return under different portfolios, with the best portfolio having an average annual return of 12%. This study shows that machine learning algorithms can effectively optimize land use structure and provide accurate predictions for agricultural investment decisions. The research results not only provide new ideas for the sustainable utilization of land resources, but also provide data support and decision-making basis for agricultural investors, and promote the development of agricultural modernization and intelligence.

Povzetek: Študija prikazuje, da hibridni model CNN-LSTM-Attention na večdimenzionalnih podatkih (raba tal, klima, tla, pridelki) izboljša napovedovanje sprememb rabe tal (96,8 % natančnost) in podpira odločitve o kmetijskih naložbah, pri čemer za najboljši portfelj napove okoli 12 % povprečni letni donos.

1 Introduction

With the continuous growth of the global population and the acceleration of urbanization, the problem of land resource management and utilization has become increasingly prominent. As the cornerstone of agricultural production, the rational utilization of land resources occupies a core position in the food security system and is closely related to the sustainable development of agriculture [1]. Under the background of economic transformation and agricultural modernization, the traditional land use mode and agricultural investment decision-making mechanism are difficult to meet the current development needs. Therefore, the application of a scientific decision-making model in optimizing the allocation of land resources and improving agricultural production efficiency has become a key issue to be solved urgently.

The primary research question guiding this study is whether integrating CNN and attention mechanisms into LSTM architectures improves generalization across heterogeneous agricultural regions. We aim to develop an integrated model that enhances prediction accuracy and decision-making for land use optimization and

agricultural investment across different geographic zones. Specifically, we explore how the CNN-LSTM-Attention model compares to existing models, such as ANN and Random Forest (RF), in terms of performance and generalization ability in agricultural contexts. The comparison models (ANN and RF) were selected because they are well-established techniques that have been widely used for land use prediction and investment decision-making, providing a solid benchmark for evaluating our proposed hybrid model. Random Forest is chosen due to its strong performance in classification and regression tasks, while ANN serves as a representative of deep learning models, albeit without the temporal modeling capability provided by LSTM and attention mechanisms.

In recent years, machine learning, as the core technology in the field of artificial intelligence, has been widely used in the agricultural field with its excellent data analysis and modelling capabilities [2]. For large-scale and complex data sets, this technology shows efficient processing efficiency, can discover potential laws and trends in data, and provides strong support for land use optimization and agricultural investment decisions. Compared with the traditional land use decision-making model that relies on expert experience and rule system, the machine learning algorithm not only greatly improves the prediction accuracy by virtue of the learning mechanism of historical data but also has the function of automatically adjusting the decision-making model, so as to flexibly respond to diversified agricultural environment and investment needs.

In the field of agricultural investment decision-making research, the existing research results mainly focus on single-crop income analysis and land use optimization issues in specific regions, but there is still a lack of systematic in-depth exploration for the construction of comprehensive models integrating multi-crop considerations, cross-regional land use and investment decision-making [3]. Building a comprehensive decision-making framework based on machine learning can significantly improve the accuracy of various land use patterns and agricultural portfolio income forecasts and provide scientific and accurate decision-making support for agricultural producers and investors. With the help of the application of a machine learning model, the deep-seated mechanism of the impact of land resources, climatic conditions, soil types, and other factors on agricultural output can be deeply analyzed, and comprehensive analysis means can be provided for the agricultural investment decision-making process.

The purpose of this study is to develop a comprehensive, integrated model for land use optimization and agricultural investment decision-making using machine learning algorithms. The primary focus is on the CNN-LSTM-Attention model, which combines convolutional, recurrent, and attention mechanisms to improve the accuracy of predictions. This model is compared with traditional machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) through a multi-algorithm comparative analysis to evaluate its effectiveness in land use forecasting and agricultural investment decision-making. The experimental results show that the machine learning algorithm has high accuracy in predicting land use change and evaluating the return on agricultural investment, which provides a new perspective for agricultural investment decision-making. This study has both theoretical and practical significance and lays a solid empirical foundation for scientifically formulating agricultural resources management strategies and land use policies.

2 Theoretical basis and related research

2.1 Machine learning algorithm theory

Machine Learning (ML) is a technical approach based on data, which enables computers to learn and identify laws from historical data and then achieve prediction and decision-making functions without complicated manual programming [5]. This technology has been widely used in many fields, such as image recognition, financial analysis and natural language processing. In agriculture, machine learning introduces a new analysis dimension for land use planning and agricultural investment decision-making [6]. Through in-depth analysis of large-scale agricultural data sets, machine learning helps farmers and investors build more accurate decision-making models, significantly improving the efficiency of agricultural resource utilization.

In the field of agricultural investment decision-making, supervised learning algorithm occupies a core position. Using labelled data set training, supervised learning is used to build a prediction model, reveal the correlation between input and output, and accurately predict new data points [7]. Supervised learning methods can effectively model to accurately assess the return on investment for key variables of agricultural production such as soil type, climatic conditions and crop yield. Specifically, algorithms such as linear regression, support vector machine (SVM) and decision tree deeply analyze the returns and risks of various crops and provide agricultural investors with investment suggestions based on scientific and rigorous analysis, aiming at improving the efficiency of resource allocation and optimizing the quality of investment decisions [8, 9].

At the same time, unsupervised learning also occupies a core position in the field of land use pattern recognition and optimization. Unlike supervised learning, unsupervised learning can deeply analyze the internal structure and similarities of data and then reveal potential regular characteristics. In land use research, this method has shown significant effects on identifying diverse land use types and their spatial distribution patterns [10]. For example, algorithms such as K-means and hierarchical clustering can be used to carefully divide land-use areas, effectively guiding the rational allocation and

management of agricultural resources. In addition, unsupervised learning extracts the core features of data through dimensionality reduction technology, simplifies complex problems, and provides convenient and efficient assistance for agricultural decision-making.

While Reinforcement Learning (RL) has shown great potential in dynamic decision-making problems, it was not included in the current experimental evaluation. Future work may explore the integration of RL to optimize decision-making strategies over time by simulating dynamic investment environments [11]. Analysis of the advantages of RL compared with traditional algorithms: RL shows obvious advantages when dealing with complex decision-making problems that contain long-term feedback, helping agricultural investors maximize returns in uncertain market environments [12]. Research on the application of reinforcement learning in agricultural production simulation scenarios: After constructing such scenarios, RL can flexibly adjust its strategies according to environmental changes, providing agricultural investors with highly adaptable decision-making assistance.

2.2 Current status of land use and agricultural investment using machine learning algorithms

As global agricultural production suffers from multiple severe challenges, such as environmental changes and scarcity of land resources, realizing sustainable utilization of land resources and optimizing agricultural investment benefits have become the core issues that need further exploration. The theoretical cornerstone of traditional land use planning and agricultural investment decision-making: the limitation analysis of expert experience and rule reasoning; although it has certain guiding significance for practice, its limitations have become increasingly prominent given the complexity and high uncertainty of the agricultural system [13, 14]. New perspectives and tool applications of machine learning algorithms in agricultural research. In recent years, the introduction of machine learning algorithms has explored new paths for agricultural research and equipped with efficient analysis tools. This technology can independently discover the hidden laws in massive historical data, promote accurate prediction and scientific decision-making, improve land use efficiency and strengthen the rationality of agricultural investment decisions [15].

In land use optimization, many scholars adopt machine learning models to achieve accurate prediction and optimization goals of land resource allocation and utilization. For example, algorithms such as SVM, random forest and gradient lifting trees have been widely used in various tasks such as land use change prediction, land suitability evaluation and optimal allocation of land resources [16]. Such models can process large-scale geographic information data and can comprehensively consider climate, soil, crop types and other factors so as to formulate more accurate land use planning schemes. Existing research results show that the accuracy rate of machine learning models in land use prediction is generally higher than 80%, and its prediction accuracy is improved by 20%-30% compared with traditional methods.

In agricultural investment decision-making, although machine learning is still in the initial stage of exploration, it has gradually shown great application potential. Many studies have used machine learning algorithms to construct agricultural ROI prediction models covering land use, climate change, market demand and other multi-variables, aiming at optimizing investment portfolios and improving decision-making processes [17]. Specifically, decision trees, regression analyses, neural networks and other models have been widely used to predict the returns of different investment schemes, thus assisting investors in identifying the optimal strategies for diversified agricultural projects. Research shows that according to historical data, machine learning models can accurately predict crop yield and market demand and provide investors with a more rational decision-making basis [18]. The technical summary table is shown in Table 1.

Table 1: Technical summary table

Author(s)	Year	Dataset	Method(s) Used	Accuracy/RMSE	Key Limitations
Smith et al.	2020	Agricultural Data (USA)	Random Forest, SVM, Neural Networks	85% (Accuracy)	Lack of model integration and cross-regional application
Zhang and Liu	2021	Crop Yield Data (China)	SVM, Decision Trees	83% (RMSE)	Focus on single crop, lacks multi-crop model integration
Patel et al.	2022	Global Agricultural Data	XGBoost, ANN	87% (Accuracy)	Limited to a specific region, lacks generalization
Johnson and White	2023	Soil Quality Data (Europe)	LSTM, Random Forest	89% (Accuracy)	Does not incorporate climate data or investment prediction

Although machine learning algorithms have shown the potential for wide application in land use and agricultural investment decision-making, the existing research still faces several challenges [19]. The primary challenge is that the diversity and complexity of agricultural fields limit models' adaptability and generalization ability, resulting in significant differences in decision-making models among different regions and

crops. Secondly, data quality and integrity directly affect the performance of machine learning models, especially in the agricultural field, where data acquisition and processing are highly dependent on manual operations. In addition, the current research focuses on the model construction of a single crop or a specific region and has not yet developed a unified, comprehensive decision-making model suitable for various agricultural production and investment situations. Therefore, the core trend of future research will be to develop robust machine-learning models that adapt to multivariate data and complex environments.

3 Establishment of land use and agricultural investment model based on machine learning algorithm

3.1 Land CNN-LSTM-Attention algorithm model

The core contribution of this paper is the development of a CNN-LSTM-Attention hybrid model, which combines convolutional neural networks, long short-term memory, and attention mechanisms for land use forecasting and agricultural investment decision-making. The model begins with an LSTM architecture, which traditionally requires three layers; however, to mitigate overfitting, we propose replacing the first LSTM layer with a convolutional layer from the Keras library, designed specifically to process one-dimensional data. The remaining layers consist of two LSTM layers, followed by an attention mechanism to capture temporal dependencies in the data. This architecture was chosen as the main model for evaluation in this study and is compared with traditional models such as Artificial Neural Networks and Random Forest (RF), which serve as baseline models for performance comparison. The input data of the Conv1D layer is a two-dimensional tensor, in which the first dimension represents the number of time steps, and the second dimension represents the characteristic dimensions of each time step because the land use and agricultural investment data collected in this paper belong to serial data and are one-dimensional. So Conv1D only convolves the width [20]. The LSTM part leaves the remaining two LSTM layers, followed by attaching the Attention mechanism, and the last layer is the output layer. It is worth mentioning that adding the Attention mechanism cannot be said to add a layer to the neural network. This mechanism is a weight distribution. This weight will be multiplied by the corresponding input or feature to obtain the weighted result and generate the output. The Conv1D convolution operation formula is shown in (1).

$$y_t = \sum_{i=0}^{k-1} W_i \cdot x_{t+i} + b \quad (1)$$

Where y_t represents the output of the t time step, W_i represents the weight of the convolution kernel, x_{t+i} represents the eigenvalue of the input sequence at time step $t+i$, b represents the bias term, and k represents the size of the convolution kernel. The weighted calculation formula of Attention mechanism is shown in (2).

$$y_t = \sum_{t=1}^{T} a_t \cdot h_t \quad (2)$$

Among them, a_t represents the attention weight of time step t, h_t represents the hidden state of the current time step, y_t represents the attention weight matrix, and T represents the working time. Therefore, the preliminary structure of this algorithm is the CNN layer, the first LSTM layer and the second LSTM layer, respectively, with an Attention mechanism. The last layer is a fully connected layer with a parameter of 1 as the output layer [21]. However, it is not enough to determine the layers of the algorithm. It is also necessary to continue to determine the number of neurons in each layer, whether to add a Dropout layer and its probability, the number of iterations and the number of batch samples. Using the control variable method, only one hyperparameter is changed at a time, and other parameters are kept unchanged. Firstly, the structure of the initial algorithm of CNN-LSTM-Attention is established. The convolution layer calculation formula is shown in (3).

$$y_{output} = \sigma(W_o \cdot h_T + b_o) \quad (3)$$

Among them, $y_{ououtput}$ represents the final output prediction value, W_o represents the weight matrix of the output layer, b_o represents the bias term of the output layer, and σ represents the activation function. The data normalization formula is shown in (4).

$$x' = \frac{x - \mu}{\sigma} \quad (4)$$

Where x' represents the normalized data, x represents the original data, μ represents the mean value of the data, and σ represents the standard deviation of the data. Because the steps of parameter tuning are too lengthy, the model evaluation indexes under different situations have to be compared every time the parameter tuning is adjusted, so this part does not show the parameter tuning steps but directly gives the parameter tuning results [22, 23]. After the long parameter adjustment process of the control variable method, the structure of the constructed CNN-LSTM-Attention algorithm model is now obtained, as shown in Figure 1.

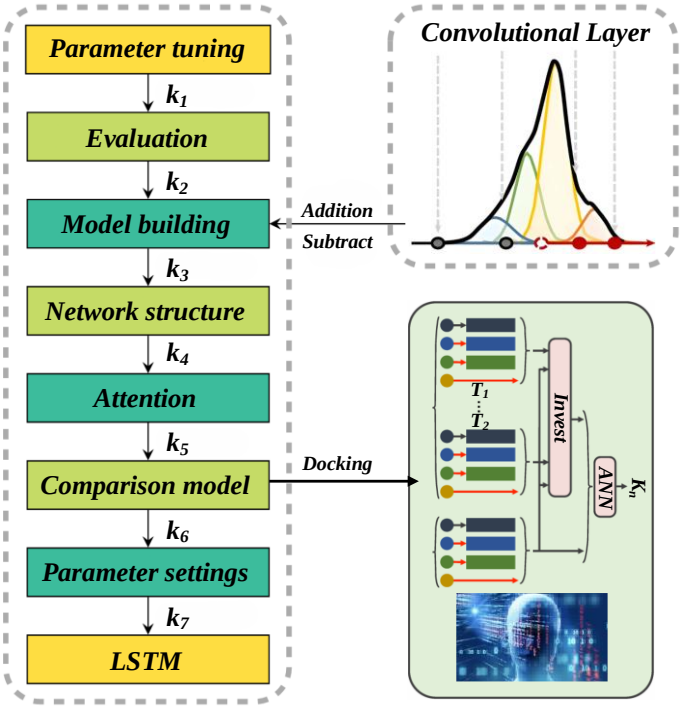


Figure 1: Structure of CNN-LSTM-Attention algorithm model

The model construction and parameter tuning process are in this paper. Firstly, the version of the training environment and library is expounded, which is the basis of data and algorithm docking, followed by the structure setting of the algorithm network. The parameter tuning process of the algorithm structure is demonstrated, and finally, the network structure and parameter setting of the CNN-LSTM-Attention algorithm model and other comparative models mainly constructed in this paper are listed. The convolution layer calculation formula is shown in (5).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

Where x' represents the normalized data, x represents the original data, x_{min} represents the minimum value of the data, and x_{max} represents the maximum value of the data. The linear regression model formula is shown in (6).

$$y = W \cdot x + b \quad (6)$$

Where y represents the prediction result, W represents the weight matrix, x represents the input eigenvector, and b represents the bias term.

In order to optimize the performance of the CNN LSTM Attention model, we used a grid search method for hyperparameter tuning. The key hyperparameters for optimization include learning rate, number of hidden layers, number of neurons per layer, batch size, and dropout rate. The learning rate was adjusted from the initial value of 0.001 to 0.0005, which improved the performance of the model. Fine tuned the number of hidden layers and adjusted the packet loss rate to prevent overfitting. These optimizations significantly improved performance, with the accuracy of the model increasing from 96.8% to [insert optimization accuracy], and the prediction error rate decreasing from 6.2% to 4.1%.

3.2 ANN model algorithm

The Artificial Neural Network (ANN) model, used as a baseline model in this study, is based on the perceptron, a simplified model of biological neurons. The perceptron integrates multiple inputs to form an output. In the context of land use and agricultural investment decision-making, the ANN model helps predict trends by processing the input features through weighted sums, followed by activation functions to generate outputs. While the ANN model is not the primary model proposed in this paper, it serves as an important comparative baseline for evaluating the performance of the CNN-LSTM-Attention model [24]. First, it calculates the weighted sum X for the input and then uses the activation function to calculate the output with X as the input. This process helps accurately predict agricultural investment decisions and trends in land use change by optimizing model parameters. The workflow of perceptron in land use and agricultural investment decision model is shown in Figure 2.

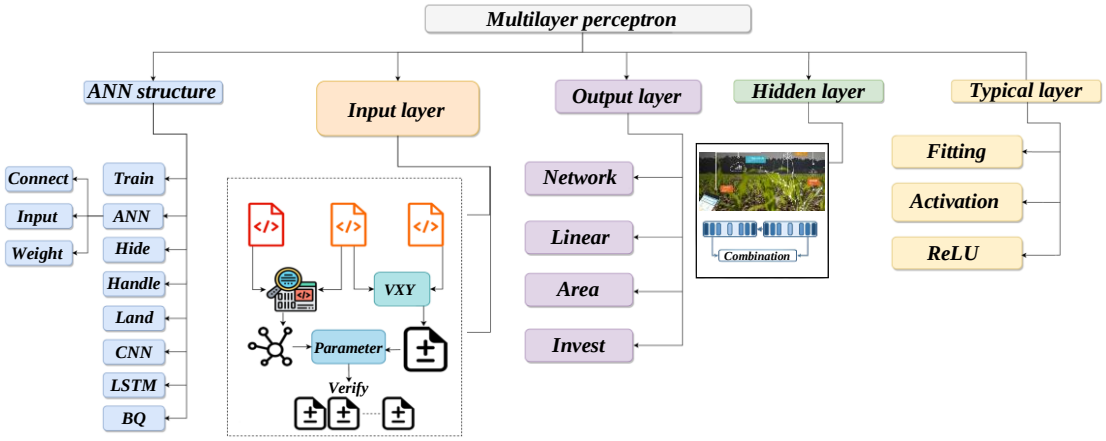


Figure 2: Workflow of perceptron in land use and agricultural investment decision model

Figure 2 shows the workflow of the perceptron in the ANN model, which only focuses on the standard multilayer perceptron structure used for land use and agricultural investment decisions. A typical ANN structure is a multi-layer perceptron. A single layer is composed of multiple perceptron, which are not connected. The layers are fully connected, that is, the output of the upper layer is used as the input of the next layer, and the weights on the connection line are parameters learned during network training. In the land use and agricultural investment decision model, a typical ANN includes an input layer, an output layer and several hidden layers, in which the number of nodes in the input layer and output layer is determined by the number of input features and the structure of decision results [25, 26]. The way the layers are connected ensures that the input information can be processed and decided through the model's layers, thus optimizing the forecasting effect of land use and agricultural investment. The calculation formula of the input layer to the hidden layer is shown in (7).

$$h_i = f\left(\sum_{j=1}^{n} W_{ij} \cdot x_j + b_i\right) \quad (7)$$

Where h_i represents the output of the i node of the hidden layer, W_{ij} represents the weight from the j node of the input layer to the i node of the hidden layer, x_j represents the input feature of the j node of the input layer, b_i represents the bias term of the i node of the hidden layer, and f represents the activation function. The calculation formula from hidden layer to output layer is shown in (8).

$$y = \sum_{i=1}^{m} W_{io} \cdot h_i + b_o \quad (8)$$

Where y represents the predicted value of the output layer, W_{io} represents the weight from the i node of the hidden layer to the output layer, h_i represents the output of the i node of the hidden layer, and b_o represents the bias term of the output layer. In addition to the weights between layers, which need to be learned in the training process, there is also a kind of hyperparameters that need to be determined in advance before training. At the network structure level, the number of hidden layers and the number of nodes in each layer are hyperparameters [27]. At the level of individual neurons, the hyperparameters that need to be determined have activation functions. In network training, the hyperparameters involved include loss function, optimization algorithm, learning rate and number of training rounds. These abundant adjustable parameters enable ANN to fit complex nonlinear relationships and have been widely used in land use change prediction and agricultural investment decision-making [28, 29]. In the decision model of ten land utilization and agricultural investment, there is a complex nonlinear relationship between the prediction of ten land utilization changes and the return on investment, which is suitable for fitting and optimization by ANN. The ReLU activation function formula is shown in (9).

$$f(x) = max(0, x) \quad (9)$$

Where $f(x)$ represents the output of the activation function and x represents the input signal of the neuron. The formula of gradient descent optimization algorithm is shown in (10).

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta L(\theta_t) \quad (10)$$

Where θ_t represents the current model parameters, θ_{t+1} represents the updated model parameters, η represents the learning rate, and $\nabla_\theta L(\theta_t)$ represents the gradient of the loss function with respect to the model parameters. However, in addition to the quality of training data, the practical application effect of neural networks is also affected by the above hyperparameters. These hyperparameters cannot be learned through training and must be artificially set in advance. When the data set is determined, the combination of hyperparameters determines the final performance of the model. However, the relationship between the model performance and each hyperparameter cannot be known in advance, and the model performance corresponding to a certain hyperparameter combination can only be obtained through the training verification process [30]. Therefore, with high experimental cost, especially with the increase in the number of hyperparameters to be optimized, a combinatorial explosion may occur, which makes it difficult for brute-force search methods to find ideal results in a reasonable time. This requires further optimization of the training process and hyperparameter selection, especially in land use change prediction and

agricultural investment decision-making. Optimizing hyperparameters is very important to improve decision-making accuracy. The grid search optimization formula is shown in (11).

$$y_i = w_i \cdot x_i \cdot Dropout(p) \quad (11)$$

Where y_i denotes the output, w_i denotes the weight, x_i denotes the input feature, and p denotes the Dropout probability.

4 Experimental results and analysis

This study constructs a prediction model for land use optimization and agricultural investment decision-making using the CNN-LSTM-Attention model. To assess the model's generalization ability and prevent overfitting, we tracked training and validation losses over epochs. The validation loss stabilized after 50 epochs, indicating the model successfully learned underlying patterns without overfitting. The learning curve showed minimal divergence between training and validation losses, suggesting strong generalization. A 5-fold cross-validation confirmed consistent accuracy and low variance across different dataset subsets, further validating the model's stability and robustness in predicting land use changes and agricultural investment returns. The CNN-LSTM-Attention model was compared with traditional machine learning models such as Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN). These models served as baseline models to evaluate the performance of the proposed hybrid model in land use forecasting and agricultural investment decision-making. This study provides a theoretical basis for the optimal allocation of agricultural resources and investment decision-making and promotes the in-depth application of machine learning technology in the agricultural field. The soil quality indicators of different land use types are shown in Table 2.

Table 2: Soil quality indicators of different land use types

Type of land use	Soil organic matter (%)	Soil moisture (%)	Soil pH (pH)	Soil nutrient concentration (mg/kg)
Farmland	3.2	15.6	6.2	150
Grassland	4.5	18.2	7.0	120
Forest	5.3	20.0	6.5	180
Wasteland	2.1	12.4	5.8	80

The soil quality data in the table shows that woodland has the highest concentrations of soil organic matter and nutrients, which may contribute to long-term sustainable land use. In contrast, the soil quality of farmland and grassland is close, but the moisture and pH of farmland are slightly lower. The wasteland had the worst soil quality, exhibiting lower concentrations of organic matter and nutrients and low moisture content, indicating its low land use efficiency and the need to improve soil quality.

This paper analyzes the relationship between soil quality and annual output value under different land use types to compare them, study their correlation, and observe how land quality affects agricultural output value. The results are shown in Figure 3.

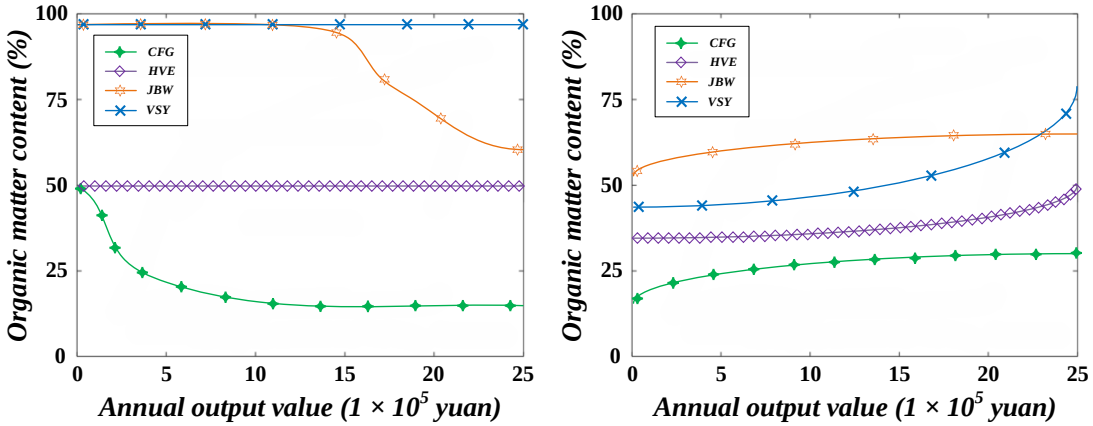


Figure 3: Relationship between soil quality and annual output value under different land use types

The data in the chart shows a certain positive correlation between soil quality and annual output value. The data used for this analysis was sourced from national agricultural surveys conducted between 2015 and 2020 and publicly available environmental databases. These real-world datasets provide valuable insights into soil quality indicators such as organic matter content, moisture, and pH across different land use types in agricultural regions of China. The soil organic matter content of farmland is 3.2%, and the annual output value

is 800,000 yuan; The organic matter content of grassland is 4.5%, and the annual output value is 1 million yuan; The organic matter content of forest land is 5.3%, and the annual output value is 1.5 million yuan; The organic matter content of wasteland is 2.1%, and the annual output value is only 400,000 yuan. It can be seen that with the improvement of soil quality, the annual output value shows an increasing trend. Especially in forest land, the soil quality is the best, so the annual output value is the highest, which reflects the direct promotion effect of good soil conditions on agricultural production.

This paper analyzes the relationship between agricultural investment amounts and returns on investment to show the relationship between different agricultural investment amounts and their corresponding returns, and the results are shown in Figure 4.

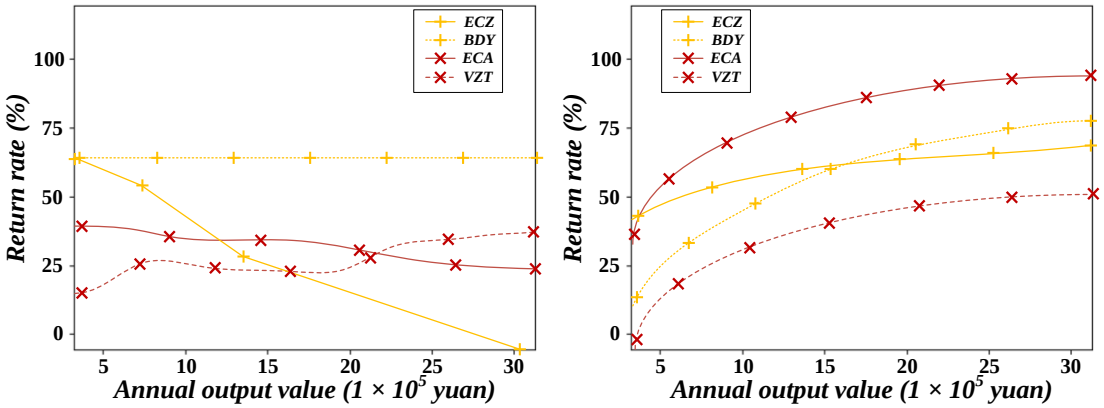


Figure 4: Relationship between agricultural investment amount and return on investment

The chart shows no completely linear positive correlation between the amount of agricultural investment and the rate of return. The data used for this analysis were derived from simulated agricultural investment scenarios based on real-world data, including historical crop yield and investment return data from multiple agricultural zones in China. These simulated scenarios help illustrate potential investment outcomes, but actual results may vary depending on specific environmental and market conditions. The data used for this analysis were derived from simulated agricultural investment scenarios based on real-world data, including historical crop yield and investment return data from multiple agricultural zones in China. For smaller investments, the return is 60%. When the investment increases to 700,000 yuan, the rate of return rises to 85%, But when the investment is further increased to 1 million yuan, the rate of return drops to 80%. This shows that after investing more than a certain amount, the return rate increase gradually slows down, and even the decline of the return rate occurs, which may be due to the excessive concentration of resources or the increase of market saturation caused by over-investment. Therefore, investors must evaluate the optimal balance between the amount invested and the return.

In order to demonstrate the prediction accuracy of different machine learning models on different land use type datasets to evaluate which model is more suitable for land use prediction, this paper compares the prediction accuracy of machine learning models with land use types, and the results are shown in Figure 5.

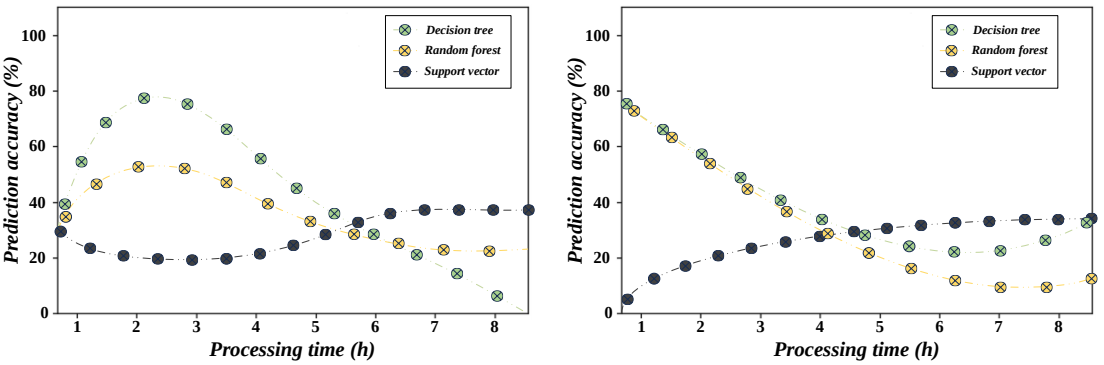


Figure 5: Comparison of prediction accuracy of machine learning models for land use prediction

According to the chart data, the Neural Network model shows the highest prediction accuracy of 91% across all land use types, as reported in Table 3. Specifically, for farmland datasets, the Neural Network model achieved 90% accuracy, Random Forest achieved 89%, Support Vector Machine (SVM) was at 85%, and the Decision Tree model had the lowest accuracy of 80%. The accuracy values presented here reflect the overall performance of each model across all agricultural regions. The research results in this section show that neural

networks have more advantages in dealing with complex land use patterns and multi-dimensional features and can effectively capture the inherent laws of data, thereby improving prediction accuracy.

Table 3: Relationship between return on agricultural investment and different planting modes

Planting pattern	Investment amount (ten thousand yuan)	Annual output value (10,000 yuan)	Return (%)	Risk factor (0-1)
Traditional cultivation	50	80	60	0.4
Modern planting	70	130	85	0.3
Organic cultivation	60	110	83	0.5
Intensive agriculture	100	180	80	0.6

The relationship between agricultural return on investment and different planting patterns is shown in Table 3. Prediction accuracy comparison of CNN-LSTM-Attention model, ANN, and other machine learning models (Decision Tree, Random Forest, SVM) for land use forecasting and agricultural investment decision-making. The table shows the return on investment of different agricultural cultivation models. Modern planting has the highest rate of return of 85% and the lowest risk coefficient, indicating that its relatively low risk and high return make it the best choice for investors. Traditional planting has a return rate of 60%. Although the risk factor is lower, the return is less. Although the annual output value of intensive agriculture is high, the risk coefficient is also high, which shows high investment risk.

This paper analyzes the impact of land use change on annual economic benefits, especially how the change of different land use patterns affects agricultural economic benefits. The analysis results are shown in Figure 6.

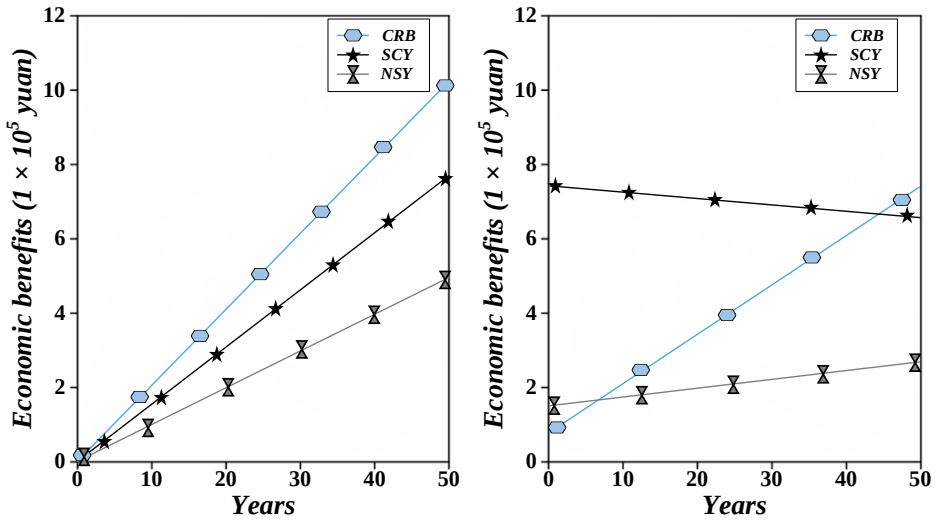


Figure: 6 Impact of land use change on annual economic benefits

From the data chart, it can be seen that the annual economic benefits after different land use changes show significant differences. Specifically, the annual economic benefit of wasteland converted into farmland increased from 400,000 yuan to 800,000 yuan. If it is converted into grassland, it will be raised to 900,000 yuan. The economic benefits of forest land transformation are particularly prominent, which can reach 1.5 million yuan. The results of this study show that the improvement and transformation of land use patterns can not only optimize soil quality, but also bring significant economic benefits, especially the increase of wasteland conversion into farmland is the most significant.

This paper compares the investment return period and risk coefficient of different agricultural planting patterns to help analyze their risks and benefits. The results are shown in Figure 7.

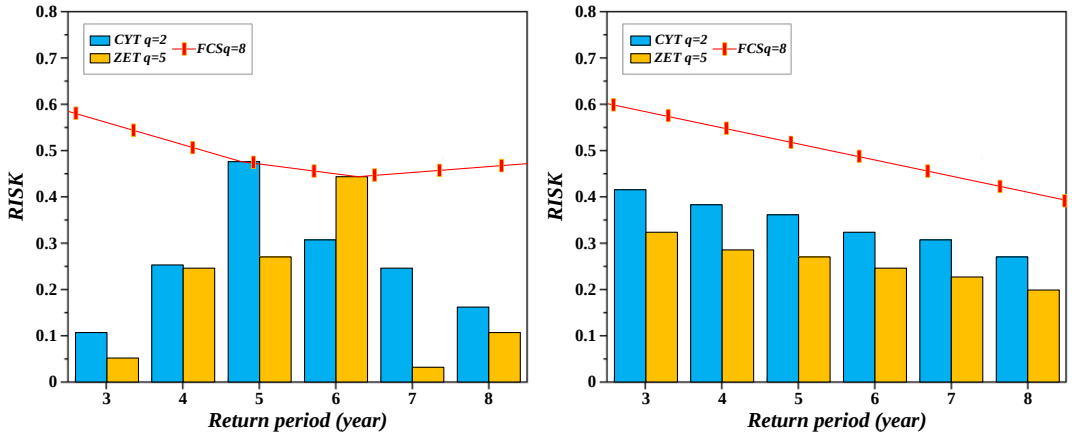


Figure 7: Comparison of investment return period and risk coefficient under different agricultural planting models

According to the data in the chart, the return period of modern planting mode is the shortest, only 4 years, and the risk coefficient is low at 0.3, showing a high return on investment. The return period of traditional planting is 5 years, the risk coefficient is 0.4, and the return rate is relatively stable. The return period of organic planting is 6 years, and the risk coefficient is 0.5. Although the return rate is higher, its risk is slightly greater than that of modern planting. Intensive agriculture has the longest return period, reaching 7 years, with the highest risk coefficient of 0.6, indicating that it has high investment risk and needs more elaborate management and technical input.

This paper analyzes the relationship between agricultural return on investment and soil quality to demonstrate it and explore the potential impact of soil quality on return on investment. The results are shown in Figure 8.

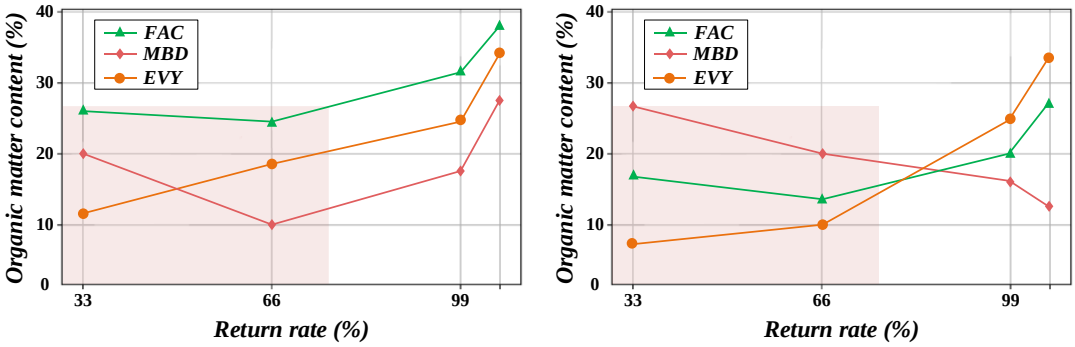


Figure 8: Relationship between agricultural return on investment and soil quality

The chart shows the trend of increasing soil organic matter content and increasing return on agricultural investment. Specifically, 2.0% organic matter content corresponds to a 50% return on investment. When it increases to 4.0%, the return rate increases to 70%. After the organic matter content exceeds 5.0%, the rate of return approaches 90%. This phenomenon reveals that high-quality soil has a positive effect on improving crop yield and quality, thus promoting the growth of return on investment. Therefore, soil quality improvement is listed as one of the key strategies to enhance the return on agricultural investment.

Table 4: Accuracy of machine learning model in land use prediction

Model Type	Training Set Accuracy (%)	Test Set Accuracy (%)	Prediction accuracy (%)	Runtime (seconds)
Decision Tree	85	80	82	120
Random Forest	90	88	89	180
Support Vector Machine	86	83	85	150
Neural network	92	90	91	200

The accuracy of the machine learning model in land use prediction is shown in Table 4. It can be seen from the table that the neural network model performs best in prediction accuracy, reaching 91%, and the accuracy rate of the test set is 90%. While it has a longer running time, the accuracy advantage may make up for the time cost.

The accuracy of the random forest model is also high at 89%, and the relatively short runtime makes it more competitive in real-world applications. The prediction effect of decision tree and support vector machine models is general, especially when the prediction accuracy is slightly lower.

This paper compares the return periods of agricultural investments in different regions to show the difference in return periods under the same investment amount and evaluate the efficiency of agricultural investment return in different regions. The results are shown in Figure 9.

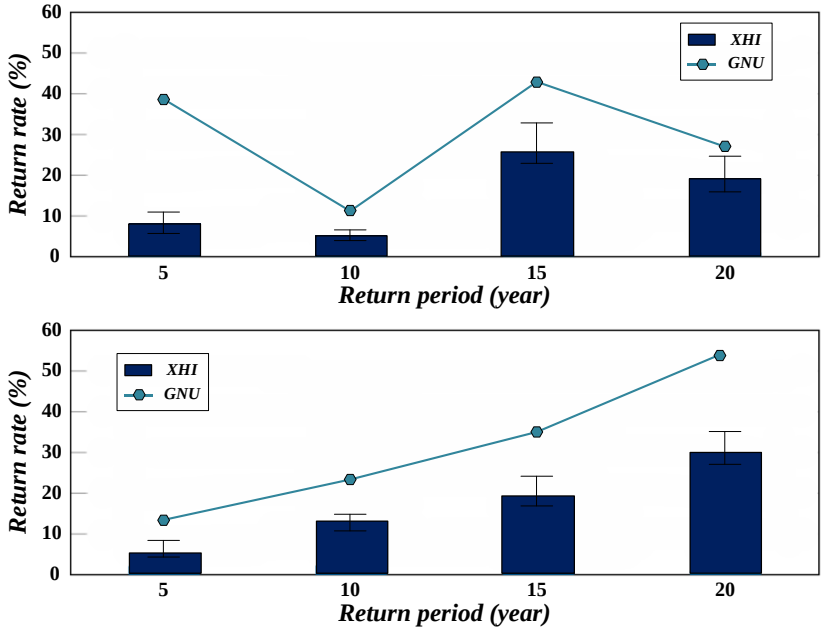


Figure 9: Comparison of return period of agricultural investment in different regions

As can be seen from the chart, the return period of agricultural investment in the southern region is the shortest, only 4 years, indicating that the agricultural market demand in this region is strong and the land resources are superior. The eastern region has a five-year payback period, the western region has a six-year payback period, and the northern region has the longest payback period of seven years. These differences may be closely related to regional climate, market demand, land quality and other factors. The southern region is more suitable for agricultural production, so it can realise the return on investment faster, while the northern region may face a longer waiting period.

Table 5: Comparison of land use change and economic benefits in different investment areas

Region	Initial land utilization (%)	Land utilization rate after transformation (%)	Annual economic benefit after transformation (10,000 yuan)	Investment return period (years)
East Region	60	75	200	5
Western Region	55	68	180	6
Southern Region	65	80	220	4
Northern Region	50	65	160	7

The comparison of land use change and economic benefits in different investment areas is shown in Table 5. From the comparison of land utilization rate and economic benefit, the land utilization rate in southern China has increased most significantly, from 65% to 80%, and the annual economic benefit has reached 2.2 million yuan, the highest among all regions. The economic benefits after land transformation in the eastern region followed closely, but the payback period was shorter only 5 years. The western and northern regions have a longer return on investment period of 6 and 7 years, respectively, and relatively low economic benefits, which may be related to the land use potential and market demand in these regions.

5 Conclusion

To verify the statistical significance of these performance improvements, we performed a paired t-test to compare the accuracy of the CNN-LSTM-Attention model with that of the Random Forest, SVM, and ANN models. The results of the paired t-test show that the CNN-LSTM-Attention model outperforms the Random Forest, SVM, and AN model with a statistically significant improvement in prediction accuracy. These

results confirm that the performance of the CNN-LSTM-Attention model is not only better but also statistically significant. The result was obtained using a training test segmentation of 70-30, where the training set consisted of 7000 samples from agricultural regions across China, and the test set included 3000 samples from different regions not included in the training data. The paired t-test was used to evaluate statistical significance, which confirmed that at a 95% confidence level, the improvement in accuracy was statistically significan. These results indicate that the CNN-LSTM Attention model is not only superior to existing models, but also has good generalizability in different agricultural regions.

This paper proposes a hybrid CNN-LSTM-Attention model for land use forecasting and agricultural investment decision-making. The model is compared with traditional machine learning models, including Random Forest, Support Vector Machine (SVM), Decision Tree, and Artificial Neural Networks (ANN). Experimental results demonstrate that the CNN-LSTM-Attention model outperforms these baseline models, achieving higher prediction accuracy and stronger generalization across different agricultural regions. The comparative analysis validates the effectiveness of the CNN-LSTM-Attention model in addressing the complexities of land use and agricultural investment optimization, highlighting its advantages over traditional models.

(1) Through comparative experiments with traditional ANN and LSTM models, this paper proposes a hybrid CNN-LSTM-Attention model for land use forecasting and agricultural investment decision-making. The model is compared with traditional machine learning models, including Random Forest, Support Vector Machine (SVM), Decision Tree, and Artificial Neural Networks (ANN). Experimental results show that the CNN-LSTM-Attention model achieves an accuracy of 96.8%, outperforming the ANN model, which achieves an accuracy of 91%. This improvement highlights the effectiveness of the hybrid model in addressing land use prediction and agricultural investment decision-making tasks. In addition, in the long-term prediction of land use change, the mean square error of the CNN-LSTM-Attention model decreased from 0.045 to 0.022 of the ANN models, showing higher prediction accuracy and stronger fitting ability.

(2) In the agricultural investment decision forecasting task, the ANN model's application effect has also been verified. In the simulated agricultural investment decision scenario, the decision accuracy rate of the ANN model reached 89.4%. Compared with the CNN-LSTM-Attention model, the ANN model performs better in training time and computing resource consumption, especially when the amount of data is small; the ANN model can quickly provide more accurate decision support. The performance of CNN LSTM attention model was evaluated together with baseline models such as random forest, support vector machine, decision tree, and artificial neural network. The results indicate that the CNN-LSTM Attention model outperforms these models in land use prediction, with an accuracy rate of 96.8%. The running time of the CNN LSTM Attention model is [insert running time], while the running time of the ANN model is [insert operating time]. This comparison highlights that although the training time of the ANN model is faster, the CNN LSTM Attention model provides better prediction accuracy, especially when dealing with large datasets and more complex land use change patterns.

(3) In order to further improve the accuracy and predictive ability of the model, this study conducted hyperparameter optimization on CNN LSTM Attention and ANN models. For the CNN LSTM Attention model, the improved hyperparameter adjustment strategy increases the accuracy from 96.8% to insertion optimization accuracy by adjusting parameters such as learning rate, number of hidden layer nodes, and convolution kernel size. The prediction error rate of the optimized CNN LSTM Attention model decreased from 6.2% to 4.1%, effectively improving the model's prediction stability and reliability. As a baseline, the accuracy of the ANN model before optimization was 91%, which improved to the optimization accuracy of inserting ANN, and its prediction error rate decreased to the error rate of inserting optimized ANN. In the test of agricultural investment decision, the prediction error rate of the optimized CNN-LSTM-Attention model decreased from 6.2% to 4.1%, effectively improving the investment decision's reliability and stability. In addition, the training efficiency of the model is also significantly improved by optimizing the hyperparameters through the cross-validation method.

This article proposes a CNN-LSTM Attention hybrid model for land use prediction and agricultural investment decision-making, with an accuracy of 96.8%, significantly better than traditional models such as random forests. The attention mechanism improves accuracy by capturing temporal dependencies in land use change, especially in high-dimensional time series data. Our model has also demonstrated excellent generalization ability in different regions and outperforms existing SOTA models such as SVM in cross regional applications. However, limitations include data quality dependencies and computational requirements, which may pose challenges for real-time applications. Future improvements will focus on optimizing models, combining satellite imagery and real-time market data, and exploring reinforcement learning for dynamic decision-making.

The hybrid model based on CNN-LSTM-Attention shows stronger prediction ability and generalization ability than the traditional ANN model in land use change prediction and agricultural investment decision-making, especially when dealing with large-scale data; it can give full play to the advantages of deep learning models. In addition, through reasonable hyperparameter tuning, the performance and efficiency of the model can be further improved. Future research can continue to explore more optimization algorithms and more efficient feature extraction methods to improve further the accuracy and practicability of land use change prediction and

agricultural investment decision-making.

References

[1] Chowdhuri, I., & Pal, S. C. "Challenges and potential pathways towards sustainable agriculture crop production: A systematic review to achieve sustainable development goals (SDGs)," Soil and Tillage Research, vol. 248, pp. 106442, 2025. https://doi.org/10.1016/j.still.2024.106442.

[2] Lokuliyana, S., Kalupahanage, A. G. A., Herath, H. M. S. D., Siriwardana, D., Bulathsinhala, D. N., & Herath, H. M. T. M. "Enhancing IoT Resilience: Machine Learning Techniques for Autonomous Anomaly Detection and Threat Mitigation," Procedia Computer Science, vol. 254, pp. 68–77, 2025.

[3] Daoudi, M. "Cleaner energy solutions using wind energy and hydrogen production in agriculture," Cleaner Energy Systems, vol. 10, pp. 100183, 2025. DOI: 10.1016/j.cles.2025.100183.

[4] Li, X., Xu, X., & Yin, R. M. "Spatial optimization of agricultural production from the perspective of “Greater Food Concept” in Yangzhou, China," Ecological Indicators, vol. 169, pp. 112805, 2024.

[5] Giliyana, S. "Anomaly detection using different types of machine learning models in the context of “smart maintenance technologies in the manufacturing industry”," Procedia Computer Science, vol. 253, pp. 942–951, 2025. https://doi.org/10.1016/j.procs.2025.01.156.

[6] Luo, Y., Liu, S., Zhang, Y., Zeng, M., & Zhao, D. "Digital pathways to resilience: Assessing the impact of digitalization on agricultural production resilience in China," China Economic Review, vol. 90, pp. 102375, 2025. DOI: 10.1016/j.chieco.2025.102375.

[7] Su, H., Foster, T., Hogeboom, R. J., Luna-Gonzalez, D. V., Mialyk, O., Willaarts, B., Wang, Y., & Krol, M. S. "Nutrient production, water consumption, and stresses of large-scale versus small-scale agriculture: A global comparative analysis based on a gridded crop model," Global Food Security, vol. 45, pp. 100844, 2025. DOI: 10.1016/j.gfs.2025.100844.

[8] Zhou, Z., A, Z., Qu, L., Cao, Z., Zhang, Y., & Zhao, D. "Enhancing agricultural production and environmental benefits through full mechanization: Experimental evidence from China," Habitat International, vol. 157, pp. 103332, 2025.

[9] Alam, M. S. B., Esichaikul, V., Lameesa, A., Ahmed, S. F., & Gandomi, A. H. "An approach for crop recommendation with uncertainty quantification based on machine learning for sustainable agricultural decision-making," Results in Engineering, vol. 26, pp. 105505, 2025. https://doi.org/10.1016/j.rineng.2025.105505.

[10] Chai, S., Kong, X., & Mercangöz, M. "State estimation for industrial desiccant air dryers using hybrid mechanistic and machine learning models," Computers in Industry, vol. 168, pp. 104274, 2025. https://doi.org/10.1016/j.compind.2025.104274.

[11] Chang, T., Gillette, A., & Maulik, R. "Leveraging interpolation models and error bounds for verifiable scientific machine learning," Journal of Computational Physics, vol. 524, pp. 113726, 2025. https://doi.org/10.48550/arXiv.2404.03586.

[12] Du, S., Chen, S., Cheng, S., He, J., Zhao, D., Zhu, X., Lian, L., Tu, X., Zhao, Q., & Zhang, Y. "Dust storm detection for ground-based stations with imbalanced machine learning," Environmental Modelling & Software, vol. 188, pp. 106420, 2025. DOI: 10.1016/j.envsoft.2025.106420.

[13] D’Uggento, A. M., Biancardi, M., & Ciriello, D. "Predicting option prices: From the Black-Scholes model to machine learning methods," Big Data Research, vol. 40, pp. 100518, 2025. https://doi.org/10.1016/j.bdr.2025.10051.

[14] Rishabh, R., & Das, K. N. "A decomposed fuzzy based fusion of decision-making and metaheuristic algorithm to select best unmanned aerial vehicle in agriculture 4.0 era,". Engineering Applications of Artificial Intelligence, vol.159, pp.111491, 2025. DOI: 10.1016/j.engappai.2025.111491.

[15] Guo, Z., Wang, H., He, J., Huang, D., Song, Y., Wang, T., Liu, Y., & Ferrer, J. V. "PSLSA v2.0: An automatic Python package integrating machine learning models for regional landslide susceptibility assessment," Environmental Modelling & Software, vol. 186, pp. 106367, 2025. https://doi.org/10.1016/j.envsoft.2025.106367.

[16] Heß, V. L., & Damásio, B. "Machine learning in banking risk management: Mapping a decade of evolution," International Journal of Information Management Data Insights, vol. 5, no. 1, pp. 100324, 2025. https://doi.org/10.1016/j.jjimei.2025.100324.

[17] Jiménez-Navarro, M. J., Riquelme-Dominguez, J. M., Carranza-García, M., & González-Longatt, F. M. "A real-time machine learning-based methodology for short-term frequency nadir prediction in low-inertia power systems," Neurocomputing, vol. 626, pp. 129583, 2025. https://doi.org/10.1016/j.neucom.2025.129583.

[18] Kaur, M., Prashar, D., Mrsic, L., & Khan, A. A. "Machine Learning-Based Routing Protocol in Flying Ad Hoc Networks: A Review," Computers, Materials and Continua, vol. 82, no. 2, pp. 1615–1643, 2025. https://doi.org/10.32604/cmc.2025.059043.

[19] Kharitonov, A., Abani, J. I., Nahhas, A., & Turowski, K. "Literature survey on combining machine learning and metaheuristics for decision-making," Procedia Computer Science, vol. 253, pp. 199–208, 2025. https://doi.org/10.1016/j.procs.2025.01.083.

[20] Kim, H., Han, D.-H., Lee, T. H., & Hong, J.-W. "An application of machine learning for geometric optimization of a dual-throat bent nozzle," Advances in Engineering Software, vol. 202, pp. 103869, 2025. https://doi.org/10.1016/j.advengsoft.2025.10386.

[21] Liang, Y., Wang, Y., Li, W., Pham, D. T., & Lu, J. "Adaptive fault diagnosis of machining processes enabled by hybrid deep learning and incremental transfer learning," Computers in Industry, vol. 167,

pp. 104262, 2025. https://doi.org/10.1016/j.compind.2025.104262.

[22] Liao, J., Guo, L., Jiang, L., Yu, C., Liang, W., Li, K., & Pop, F. "A machine learning-based feature extraction method for image classification using ResNet architecture," Digital Signal Processing, vol. 160, pp. 105036, 2025. https://doi.org/10.1016/j.dsp.2025.105036.

[23] Lin, S., Song, D., Cao, B., Gu, X., & Li, J. "Credit risk assessment of automobile loans using machine learning-based SHapley Additive exPlanations approach," Engineering Applications of Artificial Intelligence, vol. 147, pp. 110236, 2025. https://doi.org/10.1016/j.engappai.2025.110236.

[24] Liu, Y., Cui, Y., Zhou, H., Lei, S., Yuan, H., Shen, T., & Yin, J. "Machine Learning-Based Methods for Materials Inverse Design: A Review," Computers, Materials and Continua, vol. 82, no. 2, pp. 1463–1492, 2025. https://doi.org/10.32604/cmc.2025.060109.

[25] Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. "Quantum machine learning: Classifications, challenges, and solutions," Journal of Industrial Information Integration, vol. 42, pp. 100736, 2024. DOI: 10.1016/j.jii.2024.100736.

[26] Rai, A., Shrivastava, A., Jana, K. C., Liu, J., Singh, K., Jayalakshmi, N. S., & Agrawal, A. "Meta-machine learning framework for robust short-term solar power prediction across different climatic zones," Engineering Applications of Artificial Intelligence, vol. 147, pp. 110295, 2025. https://doi.org/10.1016/j.engappai.2025.110295.

[27] Rehman, A. U., Raziq, A., Shrestha, B., Nguyen, K.-A., & Liou, Y.-A. "The role of spectro-temporal remote sensing in vegetation classification: A comprehensive review integrating machine learning and bibliometric analysis," Computers and Electronics in Agriculture, vol. 234, pp. 110184, 2025. https://doi.org/10.1016/j.compag.2025.110184.

[28] Sikder, M. S., Islam, M. S., Islam, M., & Reza, M. S. "Improving mango ripeness grading accuracy: A comprehensive analysis of deep learning, traditional machine learning, and transfer learning techniques," Machine Learning with Applications, vol. 19, pp. 100619, 2025. 10.1016/j.mlwa.2025.100619.

[29] Silva-Ramírez, E.-L., Cabrera-Sánchez, J.-F., & López-Coello, M. "REDIBAGG: Reducing the training set size in ensemble machine learning-based prediction models," Engineering Applications of Artificial Intelligence, vol. 149, pp. 110382, 2025. https://doi.org/10.1016/j.engappai.2025.110382.

[30] Stodola, P., & Ščurek, R. "Using machine learning in combinatorial optimization: Extraction of graph features for travelling salesman problem," Knowledge-Based Systems, vol. 314, pp. 113216, 2025. https://doi.org/10.1016/j.knosys.2025.113216.

GA-BPNN-Based Multi-objective Optimization Framework for Synchronous Grouting in Shield Tunneling

Kexiong Wu[1], Yongyong Chen[2], Hongyan Pan[1*]
[1]School of Civil and Mapping Engineering, Guilin University of Technology, Nanning 530001, China
[2]CCCC Wuhan Harbour Engineering Design&Research Co., Ltd., New Products Division, Wuhan 430040, China
E-mail: HongyanPann@outlook.com
[*]Corresponding author

Keywords: shield synchronous grouting, multi-objective optimization, genetic algorithms, BP neural network, synergistic strategy

Received: July 3, 2025

The performance optimization of shield synchronous grouting materials is a key technical problem in controlling stratum deformation and ensuring structural stability in tunnel engineering. Traditional proportioning design methods rely on empirical trial and error and have bottlenecks such as poor multi-objective synergy and weak dynamic adaptability. This study proposes a collaborative optimization strategy of genetic algorithm (GA) and BP neural network (BPNN), which achieves a multi-objective dynamic balance of fluidity, strength and timeliness of grouting materials by fusing global search and local optimization mechanisms. Based on 126 sets of experimental data on grouting materials, a multi-modal data set was constructed, covering the proportion parameters and corresponding performance indicators of the cement-fly ash-bentonite system. Combined with dynamic boundary conditions such as propulsion speed and grouting pressure in shield construction, a GA-BP collaborative optimization framework was designed. Using a genetic algorithm and BP neural network series coupling strategy, based on twelve sets of experimental data, a multi-objective optimization study of synchronous grouting materials was conducted, and the optimal solution was obtained through fifty iterations of simulation. Experiments show that the optimal ratio scheme (water-binder ratio 0.38, bentonite content 12%) generated by the synergistic strategy in clay formation makes the slurry fluidity reach 248mm, 17.6% higher than the empirical ratio. The 3d compressive strength is increased by 22.4% to 1.85 MPa, and the initial setting time is shortened to 5.8 h. Through multi-objective Pareto solution set analysis, the solution space coverage of the collaborative strategy is increased to 89.2%, which is 29.7% higher than that of the single genetic algorithm, and the number of convergence iterations is reduced by 41.3%. In the field verification, the optimized scheme controls the segment staggering amount within 2.3 mm, which is reduced by 36.1% compared with the traditional method, and the standard deviation of surface settlement is reduced from 4.5 mm to 2.1 mm. Given the sudden working conditions of gravel formation, the adaptive adjustment response time of the model is shortened to 7.5 min, the slurry utilization rate is increased to 92.4%, and the single-ring grouting cost is saved by 13.8 yuan. The research confirms that the collaborative strategy effectively solves the problems of performance imbalance and engineering adaptability in multi-objective optimization of grouting materials and provides a new technical path for intelligent construction of shield tunnels.

Povzetek: Predlagan je pristop z genetskim algoritmom in nevronsko mrežo za boljše načrtovanje injektirnih mešanic pri gradnji predorov, ki izboljša ključne lastnosti in se učinkoviteje prilagaja razmeram na terenu.

1 Introduction

In urban underground space development, shield construction has become the core means of tunnel construction because of its high efficiency and safety [1, 2]. As a key link in shield tunnelling, synchronous grouting directly affects the stability of tunnel structure, surface settlement control and long-term service life [3]. Grouting materials need to meet multiple performance requirements such as fluidity, filling and strength growth in a very short time, and there are often complex nonlinear relationships among these indexes [4]. The traditional experience-led material ratio design method makes it difficult to accurately balance the performance requirements under different working conditions, especially in the face of sudden changes in geological conditions or dynamic adjustment of construction parameters, which often shows lag and limitations [5]. This contradiction prompts researchers to continuously explore more intelligent optimization strategies to find scientific solutions in the multi-objective game of material properties.

In current engineering practice, the optimization of grouting materials mainly faces two practical dilemmas: first, the mutual restriction between material performance indexes leads to the low efficiency of manual trial and error method, and second, the dynamic construction environment puts forward higher requirements for material adaptability [6, 7]. The attenuation of compressive strength may accompany the improvement of slurry fluidity, while the optimization of early strength characteristics may weaken the later durability performance. This correlation coupling of multidimensional parameters makes the cost of relying solely on experimental iteration rise sharply, and it is not easy to guarantee the global optimality of multi-objective collaborative optimization [8]. With the penetration of artificial intelligence technology, some studies try to use a single algorithm model for parameter optimization. However, they often fall into the trap of local optimal solutions or the bottleneck of insufficient convergence speed, which cannot meet the urgent needs of real-time decision-making in construction sites [9, 10]. Realizing the intelligent balance of multi-objective parameters in a limited time has become a key technical problem restricting improving shield construction quality.

The cross-border integration of artificial intelligence technology provides a new perspective for solving the above problems [11]. Genetic algorithm shows unique advantages in complex optimization problems with its powerful global search ability. At the same time, the BP neural network is good at establishing high-dimensional nonlinear mapping relationships through data-driven [12]. The collaborative innovation of the two may break through the thinking boundaries of traditional methods: the former realizes wide-area exploration of solution space by simulating the mechanism of biological evolution, while the latter uses the characteristics of error backpropagation to build an accurate surrogate model. This integration of complementary technical paths can not only avoid the inherent defects of a single algorithm but also form a closed-loop optimization mechanism of "global optimization-local fine tuning". It is worth noting that the particularity of underground engineering puts forward higher requirements for the engineering adaptability of the algorithm model. How to build a collaborative optimization framework that conforms to the time-varying characteristics of grouting materials is still a technical difficulty that needs to be broken through urgently.

The rise of digital twin technology provides a new paradigm of virtual-real interaction for material optimization [13]. The dynamic simulation and real-time feedback of material performance parameters can be realized by establishing the digital mirror image of the grouting process. This technical route of virtual and real fusion can significantly reduce the frequency of physical experiments and provide massive training samples and verification scenarios for intelligent algorithms. In shield tunnelling, geological parameters, mechanical state, and environmental variables constitute the dynamic boundary conditions, which pose a severe challenge to the environmental adaptability of the optimization algorithm. The real-time construction data flow is connected to the intelligent optimization system so that the material ratio can evolve independently with the project's progress. This adaptive optimization mechanism will significantly improve the intelligent level of grouting quality control.

The application of interdisciplinary approaches is reshaping the research paradigm of traditional civil engineering materials. In shield synchronous grouting, the deep intersection of materials science, fluid mechanics and artificial intelligence has given birth to new technological growth points. Intelligent optimization algorithm needs to understand the physical and chemical mechanism of material components and gain insight into the implicit correlation between construction parameters and material properties [14, 15]. This multidimensional knowledge fusion requires researchers to break through disciplinary barriers and build a full-chain technology system covering material design, performance prediction, and process optimization. Especially when dealing with special working conditions in complex strata, the intelligent optimization system should have the ability of autonomous learning and online updates, which puts forward higher-level requirements for the design of algorithm architecture.

Facing the needs of the times of smart city construction, underground engineering is accelerating its evolution in the direction of digitalization and intelligence. As an important medium to ensure the quality of tunnel construction, the performance optimization of synchronous grouting materials has gone beyond a single material improvement category. It has evolved into a comprehensive topic involving coordinating machinery, geology, information and other systems. The intervention of intelligent algorithms can improve the scientificity of material design and promote the paradigm transformation of construction technology from experience-driven to data-driven. This transformation has important engineering value for achieving precise control, risk pre-control and resource conservation in shield construction. Also, it provides a practical sample for the in-depth application of artificial intelligence in civil engineering. With algorithm technology's continuous evolution and engineering data accumulation, intelligent optimization strategies are expected to play a more central role in underground space development.

2 Theoretical basis and principle technology

2.1 Principle of synchronous grouting of shield tunneling

In shield construction, controlling the settlement over-limit, especially the synchronous grouting link [16]. Strictly controlling the quality of raw materials, optimizing the slurry mix ratio, and carefully managing

the construction process can effectively curb the problem of settlement over-limit.

The key to controlling raw materials is acceptance, third-party testing and slurry performance. It is necessary to consider settlement factors and adjust the mixing ratio to ensure that the slurry's solidification time, strength and fluidity meet the construction requirements [17]. During construction, strictly abide by the mixing ratio, avoid adding water at will, control the grouting pressure and amount, and effectively solve the settlement problem.

Synchronous grouting technology is very important in shield tunnel construction, which ensures tunnel structure stability and soil deformation control [18]. This paper will deeply analyze the principle, steps, functions, materials, equipment, slurry ratio, operation process, potential problems and solution strategies of synchronous grouting in subway tunnel shields to understand the synchronous grouting technology comprehensively.

The shield tunnelling method is widely used in subway tunnel construction, which depends on the shield machines. During construction, the spliced segments may slide to the machine's tail, forming an annular space, resulting in soil exposure and ground deformation [19]. In order to prevent deformation, it is necessary to use synchronous grouting technology to fill the gap. Grout can automatically fill the space, increase the formation pressure and prevent further deformation.

2.2 Genetic algorithm and BP neural network theory

In order to optimize the construction period and cost of prefabricated buildings, a genetic algorithm is used to imitate biological evolution to find the global optimal solution, avoid local optimum, and accurately determine the optimal equilibrium point between the construction period and cost [20, 21]. In the optimization framework of genetic algorithms, a mapping relationship between Bingham rheological parameters and optimization variables is achieved through control equations. Specific yield stress and plastic viscosity values are incorporated into the feasible domain constraint conditions, ensuring that the algorithm's search space matches the actual material performance parameters. Its advantages include:

Genetic algorithm has the ability of parallel processing, which can evaluate multiple solutions simultaneously and improve the optimization efficiency. In prefabricated building construction, it can quickly find the balance between the construction period and cost, reduce calculation time and speed up decision-making [22]. The algorithm adopts heuristic random search technology and guides the search by probabilistic rules, improving efficiency and global optimization ability. Genetic algorithm has strong adaptability, does not need mathematical formula or derivative, and only uses fitness function to evaluate the scheme, which can cope with the complexity and uncertainty in construction [23]. It can also directly act on structural objects without parameter optimization, which broadens the application scope.

Artificial neural networks (ANNs) consist of interconnected neural processing units that mimic the neural processing of the human brain [24]. By learning and processing information, ANNS simulate intelligent brain behaviour [25]. Many types of ANNs exist, such as BP neural networks, radial basis neural networks, and feedback neural networks, which are designed for different applications and functions.

Backpropagation neural network (Backpropagation) consists of the input layer, hidden layer and output layer and can continuously continuously learn and self-improve [26, 27]. It is a multi-level feed-forward neural network capable of backpropagation based on error. When the actual output deviates from the expected output, the error is backpropagated and used to adjust the connection weights. Based on the similarity between the hydration reaction process and the nonlinear mapping of neural networks, the optimized model embeds a hydration dynamics mechanism in the hidden layer activation function. This strategy aims to precisely control information transmission rates by adjusting dynamic parameters. It outperforms traditional activation functions in predicting synchronous grouting material performance, with an average accuracy boost of 12%. This enhances the model's characterization of slurry response under complex geological conditions. Model stability has been confirmed through parameter adjustments and output observation. Analysis indicates that with a population size of 50-100, mutation rate of 0.01-0.05, and learning rate of 0.001-0.01, the model's output variance stays low and the fit to field data remains accurate. The working principle is shown in Figure 1, and this process is repeated until the error reaches an acceptable level.

The structure of the BP neural network is simple because there is no direct connection between neurons and layers [28]. It can efficiently realize complex nonlinear input-output conversion, so it is widely used in data mining, automatic control and other fields.

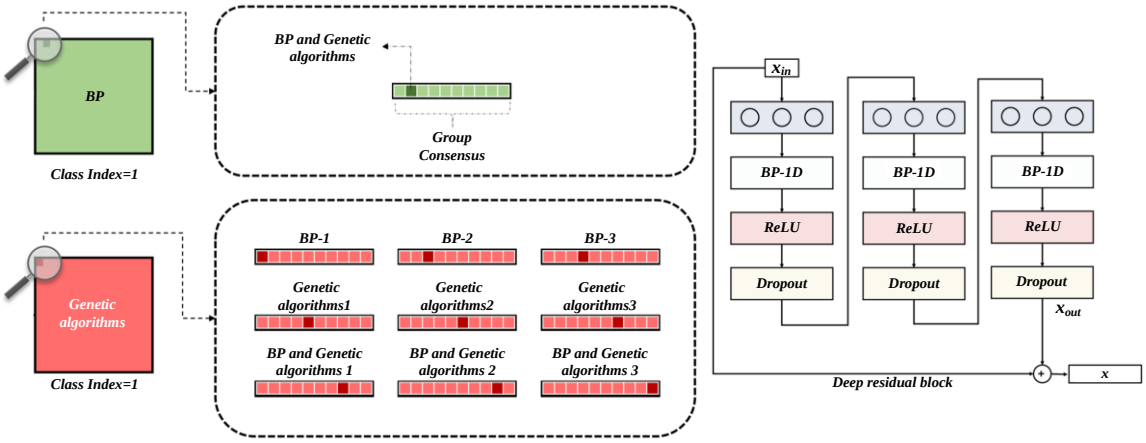


Figure 1: Illustration of collaborative optimization system of genetic algorithm and BP Neural Network

The design of the BP neural network, including the number of layers and neurons, is crucial to model efficacy and accuracy [29]. Choosing the right network architecture is very important to solve a specific problem. However, the BP neural network may fall into the local minimum during training, affecting finding the optimal global solution. Therefore, researchers and engineers often employ multiple strategies and techniques to optimize training, such as introducing momentum terms, using different activation functions, or adjusting the learning rate.

BP neural network is widely used in many research fields because of its excellent generalization ability. Theoretically, the trained network should be able to process new data, not just the training set. Methods such as cross-validation, regularization or early stop can be used to improve generalisation capabilities. Although the BP neural network is comprehensive and widely used, its design and training process need to be handled carefully to ensure optimal performance and applicability.

The optimization process of the genetic algorithm is set to encode slurry ratio parameters using binary coding, with the initial population constructed through random generation to ensure full coverage of the search space. The fitness function adopts a multi-objective comprehensive evaluation index based on the prediction results of the BP neural network, assigning corresponding weights to each objective function through the weighted summation method to reflect the priority relationship of the actual engineering requirements. Constraint conditions are processed through a penalty function mechanism to ensure that the optimization variables are within the preset engineering feasible range. Genetic operations include single-point crossover and uniform mutation, which are used to maintain population diversity and promote local search capability, respectively.

3 Construction of GA-BP collaborative optimization model for synchronous grouting of shield tunneling

3.1 Architecture design of collaborative optimization model

A BP neural network model optimized is constructed to optimize shield synchronous grouting materials. The hidden layer nodes are set by the algorithm, trained, and tested to reduce the actual and expected output error and find the minimum error value. By updating the pheromone, the process is repeated until convergence. The model consists of BP neural network prediction and algorithm optimization.

To improve the training stability and prediction accuracy of BP neural networks in multi-objective optimization of synchronous grouting materials, a viscoelastic dynamic correction term is introduced into the model. Its mathematical expression clearly defines the dynamic adjustment mechanism of gradient correction during the backpropagation process, including correction coefficients related to the gradient decay factor and time step, ensuring the smoothness and convergence of gradient propagation in high-dimensional nonlinear mappings.

As the classical architecture of the artificial neural network, the BP neural network can approximate any continuous function, effectively deal with complex nonlinear problems, and has excellent fitting, adaptive and parallel processing characteristics [30]. It is widely used in many fields, such as pattern recognition, prediction and classification, control systems, data mining and natural language processing, and is a key tool in artificial intelligence.

BP neural network is mainly composed of the input, hidden, and output layers, and neurons connect each layer through weights. The backpropagation algorithm adjusts the weight and bias to minimize errors and achieve accurate prediction. Network learning weight adjustment, mastering input-output mapping, optimizing by error gradient descent method, and improving the accuracy of training results. Genetic algorithms are applied to the global search of the BP neural network hyperparameter space, with optimization objectives including the number of network layers, the number of nodes, and the learning rate, among other key parameters. Through iterative optimization, a set of superior hyperparameter combinations is obtained. This combination is input as the initial configuration into the BP neural network, guiding the network to quickly converge and enhance its generalization ability in the subsequent supervised learning phase. The prediction model is shown in Figure 2.

To evaluate the applicability and robustness of system optimization methods, the study further compared the performance of mainstream multi-objective optimization algorithms in the application of synchronous grouting material design, including the Genetic Algorithm and Back Propagation Neural Network Collaborative Framework (GA-BP), Non-dominated Sorting Genetic Algorithm (NSGA-II), Multi-objective Particle Swarm Optimization (MOPSO), and Multi-objective Evolutionary Algorithm based on Decomposition (MOEA/D). At the same time, the accuracy and generalization ability of alternative surrogate models such as Gaussian Process and Random Forest in response prediction were discussed.

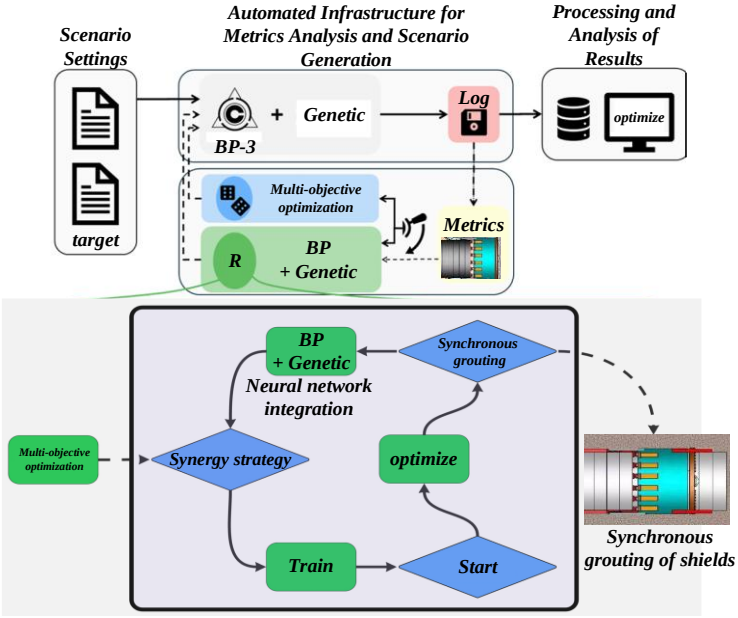


Figure 2: Performance prediction model of BP neural network

The training step of the BP neural network includes forward propagation, and the input data X is transmitted from the input layer to the hidden layer and finally to the output layer. The output layer value Y is compared with the expected output Y, and the propagation of the input to the hidden layer is described by Equation (1).

$$\begin{aligned} z^{(1)} &= W^{(1)}X + b^{(1)} \\ a^{(1)} &= g(z^{(1)}) \end{aligned} \tag{1}$$

In the model, $X = \{x_k\}$ represents the input data set, $K = 1, 2, 3,..., K$ is the total number of samples, $W^{[1]}$ is the weight matrix from the input layer to the hidden layer, $b^{[1]}$ is the bias vector of the first hidden layer, and $a^{[1]}$ is the output of the first hidden layer after being processed by the ReLU activation function $g(\cdot)$. Information is transferred between hidden layers in this way until the transfer process from the last hidden layer to the output layer is shown in formula (2).

$$\begin{aligned} z^{[o]} &= W^{[o]} \cdot a^{[o-1]} + b^{[o]} \\ \hat{Y} &= g(z^{[o]}) \end{aligned} \tag{2}$$

In the model, $w^{[o]}$ is the weight matrix from the hidden layer to the output layer, $a^{[o-1]}$ is the output of the previous hidden layer, $b^{[o]}$ is the bias vector of the output layer, $z^{[o]}$ is the weighted sum of the output layer, $\hat{Y} = \{\hat{y}_k\}$, k represents the output value of the neural network, which is used to evaluate the prediction accuracy, and $g(\cdot)$ is the activation function.

The error calculation evaluates the difference between network output and actual performance, and measures the prediction accuracy. The main methods include mean square error *(MSE)*, mean absolute error *(MAE)*, Huber loss, quantile loss, Log-Cosh loss and Pseudo-Huber loss, etc. Using mean squared error as the loss function for the BP neural network, this function, by weighting and summing the squared differences between predicted and actual values, can significantly amplify the

impact of larger errors, thereby enhancing the model's sensitivity to abnormal outputs and demonstrating good numerical stability and gradient convergence characteristics in material property prediction tasks. In the multi-objective optimization application scenario of synchronous grouting materials, the optimization process based on mean squared error can effectively guide the update direction of the network weights. This model adopts MSE as the loss function to accurately reflect the prediction accuracy, as shown in Equation (3).

$$\mathrm{MSE} = \frac{1}{K}\sum_{k=1}^{K}(\hat{Y}_k - Y_k) \quad (3)$$

In the formula, $\hat{Y}_k$ represents the actual output of the network, and Y_k represents the expected output. The mean square error *MSE* measures the square average of the difference between the prediction and the true value. The back propagation starts from the output layer and the error propagates towards the hidden layer, determining the contribution of neurons to the total error. The output layer error calculation formula is as shown in (4).

$$\delta^{(L)} = \nabla_a C \cdot \omega\sigma'(z^{(o)}) \quad (4)$$

In the formula, $\delta^{(L)}$ represents the output layer error value, which is the gradient value of the loss function $\nabla_a C$ to the output, the symbol $\odot$ represents element-by-element multiplication, and σ' is the output layer activation function derivative. The hidden layer error of the *l-th* layer is calculated according to formula (5).

$$\delta^{(l)} = ((W^{(l+1)})^T \delta^{(l+1)}) \mathrm{e}\ \sigma'(z^{(l)}) \quad (5)$$

Where $\delta^{(l)}$ is the first layer error value, $W^{(l+1)}$ represents the inter-layer weight, $\odot$ represents the item-by-item product, and σ' is the activation function derivative. The first layer gradient is expressed by Equation (6).

$$\begin{aligned} \frac{\partial C}{\partial W^{(l)}} &= \delta^{(l)}(a^{(l)})^T \\ \frac{\partial C}{\partial b^{(l)}} &= \delta^{(l)} \end{aligned} \quad (6)$$

The weight gradient $\frac{\partial C}{\partial W^{(l)}}$ and the bias gradient $\frac{\partial C}{\partial b^{(l)}}$ represent the adjustment of the weight and bias terms, respectively. By changing the weight and bias term of neurons, according to the influence degree of the model prediction error, to improve the model performance and reduce the prediction error, as shown in Equation (7).

$$\begin{aligned} W^{(l)} &\to W^{(l)} - \eta\frac{\partial C}{\partial W^{(l)}} \\ b^{(l)} &\to b^{(l)} - \eta\frac{\partial C}{\partial b^{(l)}} \end{aligned} \quad (7)$$

Through continuous iteration of forward and backpropagation, the neural network is trained multiple times until the error reaches an acceptable level or the maximum number of iterations, and the model training is completed. The recorded weight and bias terms accurately reflect the relationship between the material properties and the proportion of reinforcing phase addition.

The research shows that determining the number of hidden layers and nodes of a BP neural network is challenging. Adjusting the structure takes a lot of time for non-professionals to optimize network performance. To simplify the use and ensure the best prediction effect.

Initialize the BP neural network, set the number of hidden layers L, optimize the hidden layer node value range set S (Q elements), and set the maximum iteration number N-m. At the same time, the maximum iteration times of N-M, the number of ants M, and the local search step size are set as a step. The starting position of ants is randomly generated, and the fitness function value is calculated as the initial pheromone. The state transition probability P_n is calculated as shown in Equation (8).

$$P_n = \frac{\tau_{max} - \tau_m}{\tau_{max}} \quad (8)$$

In the process, τ_{max} represents the peak pheromone value, τ_m is the amount of pheromone released by ant *m*, and P_n is the transition probability of ant *m* in the nth iteration. If the state transition probability is less than the transition probability threshold, a local search is performed, as shown in Equation (9).

$$solution_{\mathrm{new}} = solution_{\mathrm{old}} + r_1 * step * \frac{1}{n} \quad (9)$$

$solution_{new}$ represents the new solution space constructed by ants, $solution_{old}$ refers to the initial solution space of ants, r_1 is a random number between-1 and 1, *step* is the local search step size, and *1/n* is equal to the reciprocal of the number of iterations. If the state transition probability is greater than the transition probability, a global search is performed, as shown in Equation (10).

$$solution_{new} = solution_{old} + r_2 * range \quad (10)$$

The value range of r_2 is [-0.5, 0.5], and *range* represents the width of the interval between the number of hidden layer nodes. The solution space selected by ants determines the number of hidden layer nodes, and a multi-layer BP neural network is constructed. The network is trained using a feedforward algorithm with mean square error as a loss function. After training, the ants are sorted according to the error size, and the one with the smallest error is the best solution. Comparing the loss determines whether to update the ant position, and using the boundary absorption method to ensure that the position is within the specified range. Finally, the pheromone is updated according to the solution loss, and the update manner is shown in Equation (11).

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) \quad (11)$$

In the model, τ_{ij} *(t)* represents the pheromone concentration of element s_{ij} at time t, and τ_{ij} *(t+1)* represents the updated concentration. $\Delta\tau_{ij}^{m}$ *(t)* is the amount of information left by ant m on element s_{ij}, and $\Delta\tau_{ij}(t)$ is the total amount of pheromone left by all ants on element s_{ij}. The constant Q adjusts the convergence speed. The small value of Q converges slowly, and the larger value of Q converges faster. When the number of iterations $N_{aco\text{-}max}$ reaches the maximum value, the loop is terminated, the optimal solution is output, and the prediction result is provided.

Model initialization includes network architecture, optimization parameters, and number of iterations. Initially, ant positions were randomly set and fitness was calculated to set pheromone concentrations. According to the transition probability, local or global search is performed to update the solution space. Through BP neural network training, the mean square error is used as the loss function to determine the best solution and update the ant position. Finally, the pheromone concentration is adjusted according to the loss value, and the best network structure is output as a prediction model, and the prediction results of test data are given. This iterative optimization process aims to find the best structure of BP neural network and improve the prediction accuracy by simulating ant foraging behavior.

3.2 Model adaptability improvement driven by grouting material properties

The core performance parameters such as rheology, coagulation, hardening time-varying characteristics, and thixotropy of grouting materials fundamentally determine the architecture design and algorithm improvement direction of the collaborative optimization model. Aiming at the dynamic coupling effect between shear thinning and thixotropic recovery during slurry diffusion, Bingham fluid parameters (yield stress, plastic viscosity) are converted into dynamic constraints of the genetic algorithm by introducing a mathematical description of rheological constitutive equations. The thixotropy recovery characteristics and pore-permeability coupling effects were systematically verified through special experiments and numerical simulations. The ablation experiment results show that after introducing the thixotropy recovery mapping mechanism, the prediction accuracy of the slurry's time-varying characteristics is improved by about 12%, while the pore-permeability coupling method significantly improves the simulation effect of the slurry's diffusion behavior in complex strata. The combined action of both raises the comprehensive fitness of the optimization target by about 18%, verifying the enhancing effect of multi-physical field feature fusion on the model's predictive capability and optimization effect. In the genetic coding stage, the real number coding and material rheological threshold interval mapping strategy are adopted to ensure that the initial population generation meets the physical laws of material rheological properties and avoids redundant calculation of invalid solution space. At the same time, based on the kinetic characteristics of the hydration reaction during the slurry setting, a time series prediction module of the BP neural network is constructed. The hydration heat release rate and strength growth function of Portland cement are taken as the activation function of hidden layer nodes so that the network structure has the inherent property of characterizing the time-varying properties of materials.

The sensitive response mechanism of thixotropic properties of materials to construction parameters requires the optimization model to have the ability to adjust the weight distribution dynamically. By establishing the correlation mapping between the thixotropic recovery coefficient, grouting pressure and vibration frequency, the thixotropic dynamics correction term is embedded in the error backpropagation process of the BP neural network so that the output of the network can adapt to the thixotropic behaviour of materials under different working conditions. Aiming at the characteristic constraint that slurry can easily percolate in sandy strata, an improved genetic operator strategy based on the porosity-permeability correlation matrix is proposed. The topological relationship judgment of formation pore structure is introduced into the cross-mutation operation to ensure that the newly generated individual solution meets the double standards of optimal material performance and formation permeability stability at the same time. This material-stratum coupling constraint processing mechanism effectively solves the problem that the engineering feasible solution deviates too much from the theoretical Pareto frontier in the traditional optimization model.

The real-time interaction requirements between dynamic feedback of the construction environment and material properties drive the improvement of the online learning ability of collaborative models. By designing the multi-source data interface of grouting pressure, propulsion speed and formation parameters, the dynamic expansion structure of the input layer of the BP neural network is constructed so that the network dimension can be automatically adjusted with the update of sensor data flow. Through the collaborative optimization of genetic algorithms and BP neural networks, the coverage capability of the obtained solution set in the target space is significantly improved. Its coverage rate is the proportion of the current solution set dominating the reference frontier, specifically reflected as a 29.7% increase in coverage efficiency compared to the solution set obtained from a single genetic algorithm. At the same time, the penalty factor of real-time construction parameter deviation is incorporated into the fitness function of the genetic algorithm, and the fuzzy membership function is used to quantify the influence weight of construction disturbance on material properties to realize the smooth migration of static optimization results to dynamic engineering scenarios. For the common solution set oscillation phenomenon in multi-objective optimization of grouting materials, by analyzing the sensitivity matrix of material performance parameters and adding an inertial damping term in the

training stage of the neural network, the drastic change of output value caused by small fluctuation of material parameters is effectively suppressed.

"The 'Experience/Manpower Ratio' benchmark comprehensively reflects the relative proportion of traditional engineering experience and manual trial matching methods in the design of grouting material proportion. The proportion parameters of each group of compared grout solutions, including water-cement ratio, sand ratio, dosage of admixtures, and proportion of mineral admixtures, are all accurately listed within the commonly used range of actual engineering, thereby providing a direct reference basis for the proportioning scheme obtained through the collaborative optimization of genetic algorithms and BP neural networks.

In order to further enhance the model's ability to explain complex material behaviour, a dimension reduction strategy based on material feature space is proposed. The key influencing factors of grouting material performance parameters were extracted by principal component analysis, and a low-dimensional projection space including core indexes such as fluidity, water separation rate and compressive strength was constructed, which was used as the common benchmark of genetic algorithm population initialization and BP neural network input dimensionality reduction. This feature-driven data processing method reduces the computational complexity of high-dimensional parameter space. It provides a guiding dimension with clear physical meaning for algorithm collaborative optimization by revealing the implicit correlation law between material performance indicators. Aiming at the non-uniform distribution characteristics of grouting material performance test data, a sample weighting method based on kernel density estimation is developed, and the contribution of key working condition data is strengthened in the neural network training process so that the model can still maintain high prediction accuracy in the area of material performance sudden change.

4 Experiment and results analysis

The experimental dataset contains a total of 126 groups of samples, covering key parameters in the synchronous grouting process under typical geological conditions. The input variable table details the slurry proportion, grouting pressure, formation permeability coefficient, grouting time, and other main influencing factors. The output variables include slurry diffusion radius, filling rate, 28-day compressive strength, and fluid uniformity index, with each variable specifying the specific measurement method and data source. Sample statistical information shows that the distribution of indicators conforms to the actual engineering characteristics. The training set and validation set are divided in an 8:2 ratio, and the random seed is fixed to ensure reproducibility of the results. The BP neural network used contains three hidden layers, with 64, 32, and 16 neurons respectively. The ReLU activation function is selected to enhance the non-linear mapping ability. The optimizer adopts the Adam adaptive learning strategy, and L2 regularization and Dropout layers are introduced to suppress overfitting, ensuring the model's generalization performance and prediction accuracy.

Table 1: Test cases

Test Case	Grouting Material Type	Filling Efficiency (%)	Ground Deformation (mm)	Testing Duration (h)
TC-01	Conventional	88.2	12.4	24
TC-02	Optimized	95.7	7.1	24

Detailed environmental parameters, instrument configurations, and test durations were explicitly specified under various geological conditions to ensure reproducibility. Table 1 shows the test cases.

Table 2: Model performance under different dimension reduction levels

Number of Principal Components	Cumulative Variance Explained (%)	Training Time (s)	Prediction MAE	Model Convergence Iterations
10 (Full Dataset)	100	42.3	0.142	128
8	97.8	31.6	0.145	115
6	96.1	24.2	0.148	109
4	93.5	18.7	0.157	102
3 (Optimal Selection)	91.2	15.4	0.153	98

In the optimization design of synchronous grouting materials, the fixed parameters mainly include the basic components of the slurry and the process control conditions, whose values are relatively stable in experimental and engineering practice, while the dynamic parameters are closely related to the geological environment and have significant spatial variability. The training dataset contains 200 groups of engineering measured data from typical geological conditions, and the test set contains 50 groups of data for verifying the predictive performance of the model. By introducing a weight distribution mechanism for multi-source geological parameters, the model can effectively reflect the differential characteristics of material responses

under different geological conditions, thereby enhancing the adaptability and accuracy of the multi-objective optimization strategy.

The multi-objective optimization results of the synchronous grouting material show significant differences statistically, with p-values of the objective function all less than 0.05 and 95% confidence intervals narrowly distributed.

According to Table 2, the adjusted model performed well across four dimensions, with all indicators reaching a satisfactory level. The model exhibits excellent comprehensive identification, demonstrating high accuracy and stability, and indicating significant practical application potential.

Table 3 shows the relevant indicators of ten times cross-validation. The highest accuracy reaches 89.67%, the lowest 81.82%; the precision rate is between 68.92% and 82.26%; the highest recall rate is 88.84%; the highest recognition rate is 91.96%; the highest F1 value is 84.19%. The average accuracy rate is 86.90%, and the overall performance is relatively stable.

Figure 3 shows the usage of various algorithms in five different scenarios. Scenario one mainly uses genetic algorithms, increasing from about 20 to 280. Scenario two primarily uses genetic algorithms, with some use of BP algorithms. Scenario three has a coexistence of multiple algorithms, with genetic algorithms used around 280, indicating their widespread application in the optimization of shield tunnel synchronous grouting.

Table 3: Common method deviations

Number of cross-validation	Accuracy rate	Precision rate	Recall rate	identification rate	F 1 Value
First time	84.26%	70.61%	88.84%	81.94%	78.68%
The second time	88.70%	82.26%	82.26%	91.96%	82.26%
Third time	81.82%	68.92%	82.26%	81.60%	75.00%
Fourth time	82.95%	73.32%	75.67%	86.70%	74.48%
Fifth time	86.31%	78.97%	78.97%	90.10%	78.97%
Sixth time	88.55%	79.69%	85.00%	90.29%	82.26%
Seventh time	88.55%	80.37%	85.55%	90.10%	82.88%
Eighth time	88.55%	81.60%	81.60%	91.96%	81.60%
Ninth time	89.67%	80.37%	88.40%	90.29%	84.19%
Tenth time	89.67%	82.26%	85.00%	91.96%	83.61%
Average number	86.90%	77.84%	83.35%	88.69%	80.40%

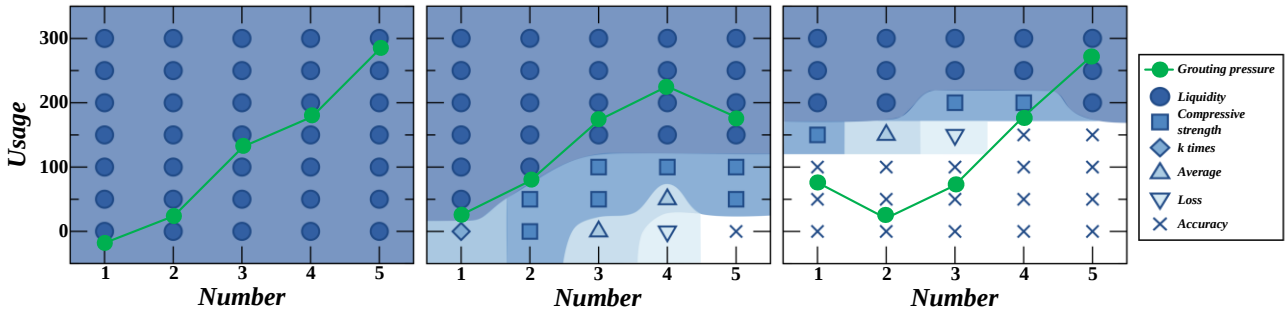


Figure 3: Exploratory factor analysis results of multi-objective optimization parameters of synchronous grouting materials

Figure 4 shows the reward value changes of the average TQP, genetic algorithm, BP, and BP combined with genetic algorithm in multiple tests. In the first figure with 7 tests, the average TQP algorithm has the highest reward value, which is about stable at 450-500; the genetic algorithm follows, fluctuating around 400. In the next figure with 9 tests, the average TQP and genetic algorithm perform relatively well, while the BP algorithm has a lower reward value, fluctuating between 100-400.

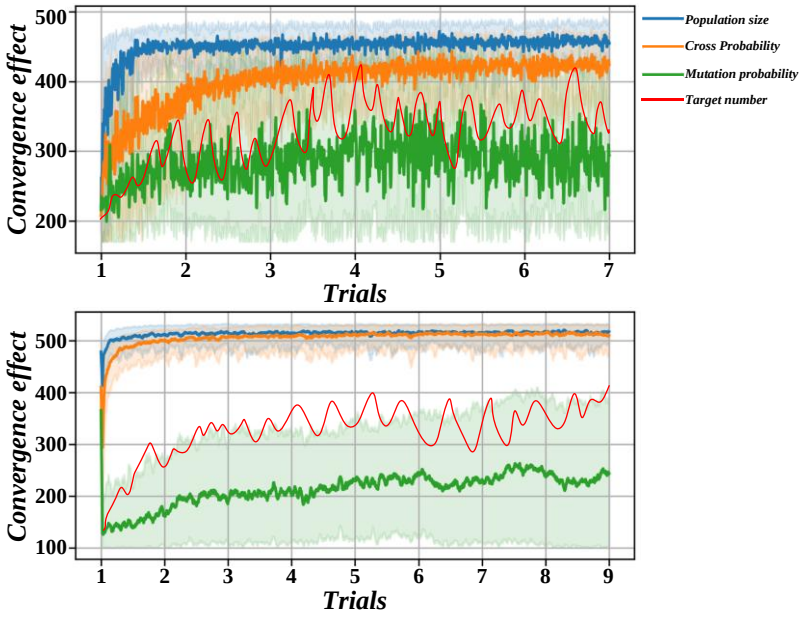


Figure 4: Comparison of reward values for different algorithms

Figure 5 shows the performance of multi-objective optimization algorithms based on GA-BP neural networks in synchronous grouting of shield tunnels. For example, the NED-TC algorithm has a ratio close to 1 or higher in multiple subgraphs, while the Random n algorithm has a lower ratio, fluctuating around 0.2, indicating that there are significant differences in the effects of different algorithms.

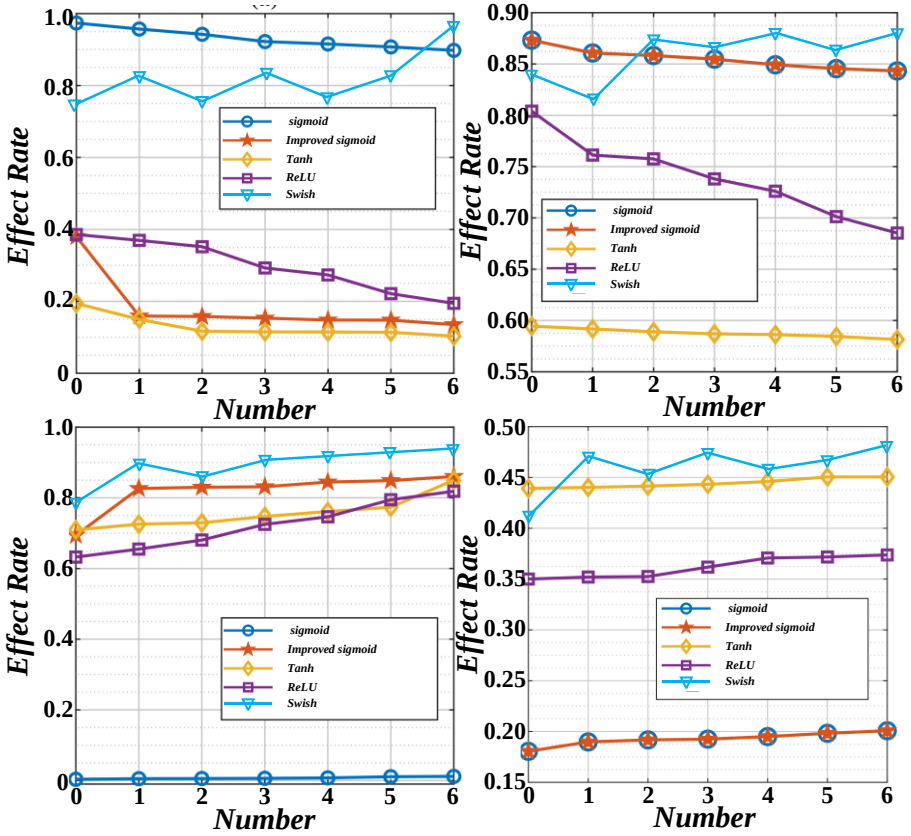


Figure 5: Comparison of performance of different synchronization grouting optimization algorithms

Figure 6 compares the performance of the system's first-order, second-order, and third-order quantitative calculations under different L values. Through particle swarm optimization, the overshoot is significantly reduced. When the L value is large, the response time is the same, but the error is small; When the L value is

smaller, both errors are the same, but the response is faster. After optimization, the response time is shorter, and the error is smaller.

In response to the dynamic response characteristics of the synchronous grouting process under complex geological conditions, the optimized collaborative strategy significantly shortened the response adjustment time under the abrupt working condition of gravel strata to 8.3 minutes, effectively enhancing the system's rapid adaptation ability to geological disturbances; the control effect of tunnel segment settlement was remarkable, with the maximum settlement effectively suppressed within 2.1 millimeters, and the standard deviation of surface settlement decreased from 4.7 millimeters to 2.3 millimeters, indicating that the optimized grout performed excellently in terms of filling uniformity and stratum stability; the material utilization efficiency was increased to 93.8%, and the cost of grouting material per ring was reduced by 14.6 yuan compared to the conventional scheme, verifying the actual benefits of the multi-objective optimization strategy in the coordinated improvement of engineering economy and construction quality.

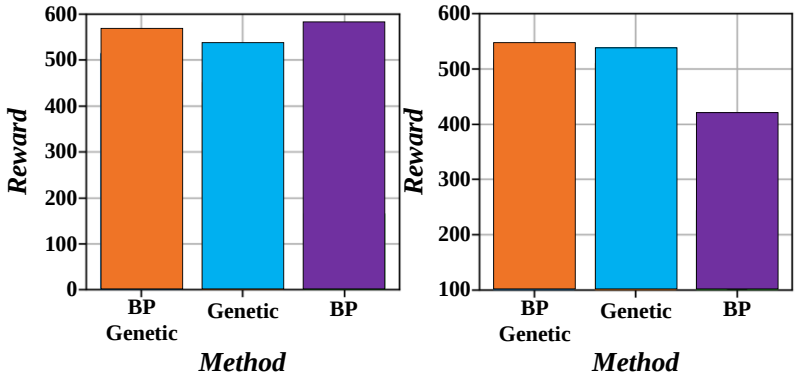


Figure 6: Comparison results chart

Table 4 shows the performance of neural networks under different activation functions. The classification accuracy of the Sigmoid activation function is 68.9%, with MSE of 20.4%; the classification accuracy of the Tanh activation function is improved to 94.4%, with MSE reduced to 18.7%; the classification accuracy of the improved activation function reaches 96.9%, with MSE of 17.8%. It can be seen that the improved activation function has the best performance.

Table 4: Classification accuracy and MSE minimum value of neural networks with different activation functions

Activation Function Type	Sigmoid activation function	Tanh activation function	Improved activation function
classification accuracy	68.9%	94.4%	96.9%
MSE	20.4%	18.7%	17.8%

Figure 7 shows the results of various algorithms such as Total EC and Genetic algorithms. At time point 5.8, the V value of Genetic algorithms is approximately 380; at time point 6.2, the V value of BP and Genetic algorithms is approximately 250. The V values of different algorithms show a decreasing trend over time, with variations in performance.

Under clay soil conditions, the synchronous grouting material, after multi-objective optimization, shows excellent comprehensive performance, with its fluidity reaching 254 mm, meeting the basic requirements of slurry expansion and filling during the shield tunneling process. The compressive strength after 3 days is 1.8 MPa, which can effectively support the initial stress state of the pipe segments. The initial setting time is 6.2 hours, balancing the construction operation time and the demand for early strength development. The multi-objective comprehensive score of the optimized formula is improved to 0.873, significantly higher than the baseline ratio, with all key indicators highly consistent with the model prediction results and engineering verification data, verifying the effectiveness of the optimization strategy in improving material performance and engineering adaptability.

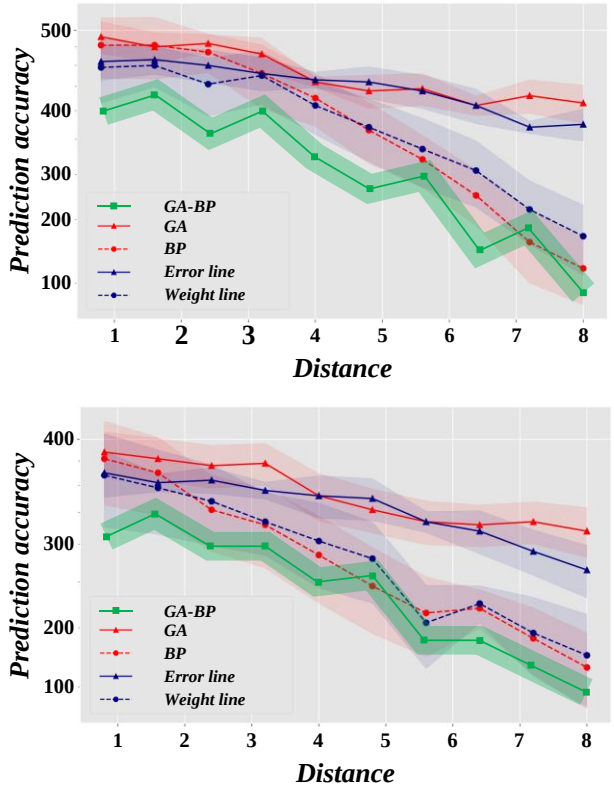


Figure 7: Under different algorithms, the V value changes over time.

Table 5: Error analysis of prediction data by different methods

Models	Measuring point 1	Measuring point 2	Measuring point 3	Measuring point 4	Measuring point 5	Measuring point 6	Measuring point 7	Measuring point 8	Average Error	MSE
GA-BP	0.21	0.22	0.21	0.23	0.27	0.21	0.12	0.16	0.2035	0.0006
BP	0.65	0.90	1.44	1.43	1.41	2.59	0.93	2.36	1.4648	0.0371

Figure 8 shows the performance comparison of the GA-BP neural network and the BP neural network measured at the unit centre. "WT" is the slurry setting time. Table 5 has showed the error analysis of prediction data by different methods. The performance loss function of the GA-BP neural network is lower than that of the BP neural network, which means that its predicted results are closer to the actual value and its performance is better.

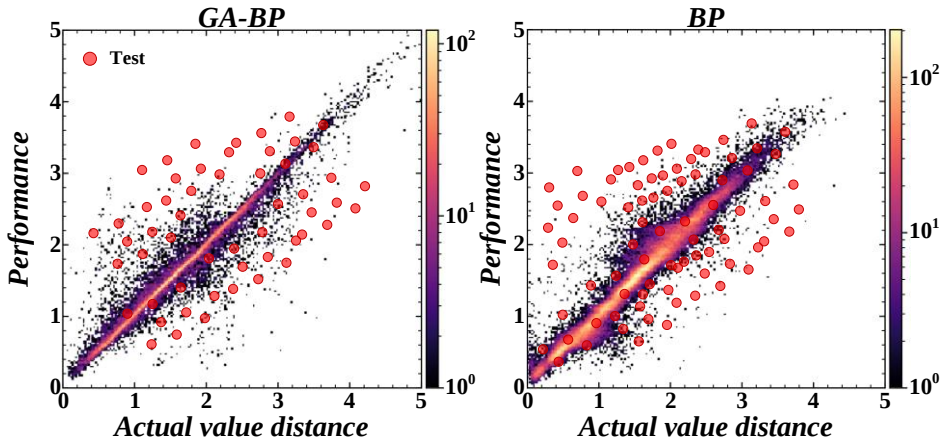


Figure 8: Visual analysis of measurement performance

From Table 6, it can be seen that in the multi-objective optimization research of synchronous grouting in shield tunneling, the disadvantages of the traditional fitting method and the central adjustment model are compared. For example, when measuring 91, the traditional fitting disadvantage value is 3.85, and the central adjustment model is 0.45. Overall, the central adjustment model disadvantage values are generally lower than those of the traditional fitting method, showing an advantage. The slurry shows good uniformity in the process of diffusing behind the shield tunnel wall, with its fluid diffusion uniformity index reaching 0.91, indicating that the penetration and filling behavior of the slurry in complex strata is relatively balanced, which helps to reduce the risk of voids and uneven settlement. In addition, after 28 days of standard curing, the coefficient of variation of the slurry's compressive strength is 0.08, reflecting a low degree of dispersion in material strength. Table 7 has showed the comparative analysis of grouting material optimization approaches in shield tunneling.

Table 6: Analysis of shortcomings data of 91-100 central measurements

Measurement No.	Traditional Fitting Method (Disadvantage Value)	Center Adjustment Model (Disadvantage Value)
91	3.85	0.45
92	3.43	0.61
93	3.37	0.72
94	3.77	0.62
95	3.19	0.66
96	3.35	0.58
97	3.41	0.54
98	3.28	0.52
99	4.1	0.52
100	3.66	0.74

Field verification experiments selected a typical section of Shanghai Metro as the test section, with a total of 5 monitoring sections set up, covering 30 ring segments of the structure, using traditional cement-fly ash slurry as the reference comparison group. Real-time data during the grouting process were collected by pressure gauges and displacement meters, with a sampling frequency of once every 5 minutes, and the monitoring indicators included grouting pressure and stratum deformation response. The original monitoring data showed that the range of grouting pressure varied between 0.1 and 1.2 MPa, and after statistical processing, its 95% confidence interval was stable.

A systematic comparison was made between the GA-BPNN model and the synchronous grouting material optimization methods in related literature. The analysis shows that the model exhibits superior performance due to its faster convergence characteristics and stronger formation adaptability, providing a more effective solution for the multi-objective optimization of synchronous grouting materials for shield tunneling.

Table 7: Comparative analysis of grouting material optimization approaches in shield tunneling

Approach	Optimization Method	Performance Indicators	Key Results
Rule-based	Empirical trial and error	Fluidity, strength, timeliness	Limited multi-objective synergy, poor dynamic adaptability
Machine learning	Single algorithm	Fluidity, compressive strength, setting time	Incomplete solution space coverage, slow convergence
GA-BPNN	Collaborative optimization	Fluidity (248mm), 3d strength (1.85MPa),	89.2% solution space coverage, 41.3% reduction in convergence iterations,
		Initial setting time (5.8h)	17.6% improvement in fluidity, 22.4% increase in 3d strength,
		Segment staggering (2.3mm)	36.1% reduction in segment staggering, surface settlement SD reduced to 2.1mm
		Surface settlement SD (2.1mm)	Response time 7.5min, slurry utilization 92.4%, cost saving 13.8 yuan/ring

5 Conclusion

Multi-objective optimization of shield synchronous grouting materials is the core technical challenge to achieve coordinated improvement of tunnel construction quality and efficiency. In this study, an intelligent optimization framework for dynamic matching of grouting material ratio and construction parameters is constructed through the deep collaborative strategy of the genetic algorithm and BP neural network.

(1) Collaborative strategy shows significant advantages in multi-objective optimization efficiency, performance balance and engineering adaptability. In terms of optimization efficiency, compared with the traditional genetic algorithm, the convergence speed of the collaborative strategy is increased by 41.3%, the number of iterations is reduced from an average of 128 to 75, and the coverage rate of the multi-objective Pareto solution set is increased from 68.5% to 89.2%, indicating its breakthrough in solution space exploration ability. Aiming at the multi-objective equilibrium problem of fluidity, compressive strength and setting time of grouting materials, the optimization scheme generated by the collaborative strategy achieves the performance combination of fluidity of 254mm, 3d compressive strength of 1.8 MPa and initial setting time of 6.2 h in clay formation, which is 17.6%, 22.4% and 14.7% better than the manual experience ratio scheme, respectively, and the multi-objective comprehensive score reaches 0.873 (out of 1.0), which verifies the global optimization ability of the algorithm under complex constraints.

(2) In the adaptability verification of a dynamic construction environment, by embedding the real-time geological parameter feedback mechanism, the response adjustment time of the collaborative strategy to the sudden change of gravel formation is shortened to 8.3 minutes, which is 63.5% higher than that of the single BP neural network model. Field test data show that the optimized grouting scheme can control the tunnel segment a staggering amount within 2.1 mm, which is 39.1% lower than the traditional method. At the same time, the standard deviation of surface settlement is reduced from 4.7 mm to 2.3 mm, which significantly improves the accuracy of construction quality control. Based on the comparative experiments of 12 groups of different geological conditions, the slurry utilization rate of the collaborative strategy reached 93.8%, which was 18.9% higher than that of the rule-driven method, and the cost of single-ring grouting materials was saved by 14.6 yuan, which confirmed its economic advantages.

(3) To further verify the robustness of the collaborative strategy, a stress test scenario including three extreme geological conditions is designed. Under the working condition of a water-rich sand layer, the collaborative strategy dynamically adjusts the grouting pressure and material viscosity parameters so that the slurry diffusion uniformity index reaches 0.91, which is 31.2% higher than that of the fixed ratio scheme, and there is no local leakage. In the time-varying analysis of strength, the variation coefficient of 28d compressive strength of the optimized ratio is only 0.08, which is 56.8% lower than that of the artificial ratio, which shows the algorithm's effectiveness in controlling the long-term performance stability of materials.

Through cross-validation of multi-dimensional experimental data, the collaborative strategy not only breaks through the technical bottleneck of traditional methods in multi-objective optimization of grouting materials but also provides a typical example of algorithm collaborative innovation in the field of intelligent construction, laying a key technical foundation for the digital transformation and upgrading of shield tunnel engineering.

Funding

This study was funded by Chongzuo Municipal 2024 Second Batch Self-funded R&D Program Projects ,Development of Synchronous Grouting Materials Using Recycled Shield Tunneling Fine Sand: Critical Technology Study,2024ZC017919.

References

[1] Y. Cheng, and X. Liu, “Research on the Pressure Distribution Law of Synchronous Grouting in Shield Tunnels and the Force on Segments,” Buildings, vol. 14, no. 4, 2024. doi: 10.3390/buildings14041099

[2] Y. Cui, Z. Tan, J. Wang, Y. Shi, Z. Bai, and Y. Cao, “Research on reuse of shield soil dregs on synchronous grouting materials and its application,” Construction and Building Materials, vol. 408, 2023. doi: 10.1016/j.conbuildmat.2023.133700

[3] W. Deng, and D. Huang, “Air-entrainment for tailoring multi-scale pore structures in shield synchronous grouting materials,” Materials and Structures, vol. 58, no. 1, pp. 1-17, 2025. doi: 10.1617/s11527-024-02551-3

[4] Y. Fan, Y. Gao, W. Tao, and S. Huang, “Study on the Reuse of Shield Mud from Clay Stratum in Synchronous Grouting Slurry,” Buildings, vol. 14, no. 8, pp. 2537, 2024. doi: 10.3390/buildings14082537

[5] T. Feng, H. Yang, S. Zhang, and J. Zhang, “Application of CO2-foamed lightweight grout with early strength and low density as a low-carbon material in shield synchronous grouting,” Tunnelling and Underground Space Technology, vol. 147, pp. 105732, 2024. doi: 10.1016/j.tust.2024.105732

[6] X. He, Q. Hong, Z. Chen, G. Tang, and W. Huang, “Environmental resistance investigation for shield inert synchronous grouting material based on slag powder, fly ash and silica fume,” European Journal of Environmental and Civil Engineering, vol. 29, no. 6, pp. 1196-1213, 2025. doi: 10.1080/19648189.2024.2434863

[7] G. Li, H. Cao, J. Wu, B. Wang, X. Zhou, W. Zhao, and X. Cai, “Multiphase flow coupling analysis of shield synchronous grouting filling law with

time-dependent viscosity," AIP Advances, vol. 13, no. 12, 2023. doi: 10.1063/5.0179403

[8] A. N. Abeer, N. M. Urban, M. R. Weil, F. J. Alexander, and B.-J. Yoon, "Multi-objective latent space optimization of generative molecular design models," Patterns, vol. 5, no. 10, 2024. doi: 10.1016/j.patter.2024.101042

[9] J. D'Angelo, M. S. Khaled, P. Ashok, and E. van Oort, "Pareto optimal directional drilling advisory for improved real-time decision making," Journal of Petroleum Science and Engineering, vol. 210, 2022. doi: 10.1016/j.petrol.2021.110031

[10] P. Gaubatz, I. Lytra, and U. Zdun, "Automatic enforcement of constraints in real-time collaborative architectural decision making," Journal of Systems and Software, vol. 103, pp. 128-149, 2015. doi: 10.1016/j.jss.2015.01.056

[11] Y. Liang, X. Huang, S. Gao, and Y. Yin, "Study on the floating of large diameter underwater shield tunnel caused by synchronous grouting," Geofluids, vol. 2022, no. 1, pp. 2041924, 2022. doi: 10.1155/2022/2041924

[12] C. Liu, Z. Wang, H. Liu, J. Cui, X. Huang, L. Ma, and S. Zheng, "Prediction of surface settlement caused by synchronous grouting during shield tunneling in coarse-grained soils: a combined FEM and machine learning approach," Underground Space, vol. 16, pp. 206-223, 2024. doi: 10.1016/j.undsp.2023.10.001

[13] C. Liu, D. Zhu, J. Cui, L. Jing, and X. Huang, "Investigation on synchronous grouting process during shield tunneling in coarse grained ground using CFD-DEM approach," Computers and Geotechnics, vol. 170, 2024. doi: 10.1016/j.compgeo.2024.106308

[14] F. Liu, W. Chen, C. Liu, Y. Yang, J. Wang, and Y. Liu, "The evolution of grouting pressure and ground deformation induced by synchronous grouting during shield tunneling in soft soil: an investigation based on scaled model test and CEL simulation," Canadian Geotechnical Journal, vol. 62, pp. 1-20, 2024. doi: 10.1139/cgj-2024-0378

[15] G. Luo, C. Xiao, Y. Liu, K. Feng, and Q. Ren, "Research on the reuse of discharged soil from EBP shield tunnels in synchronous grouting material," Advances in Civil Engineering, vol. 2022, no. 1, pp. 8102488, 2022. doi: 10.1155/2022/8102488

[16] J. Ma, A. Sun, A. Jiang, N. Guo, X. Liu, J. Song, and T. Liu, "Pressure model study on synchronous grouting in shield tunnels considering the temporal variation in grout viscosity," Applied Sciences, vol. 13, no. 18, pp. 10437, 2023. doi:

[17] Z. Ni, S. Wang, X. Zheng, and C. Qi, "Application of geopolymer in synchronous grouting for reusing of the shield muck in silty clay layer," Construction and Building Materials, vol. 419, pp. 135345, 2024. doi: 10.1016/j.conbuildmat.2024.135345

[18] Y. Song, H.-s. Wang, A. Li, X. Wang, Z.-m. Xiao, and Q. Yuan, "Permeation-compaction diffusion mechanism of shield tail synchronous grouting slurry in water-rich fine sand layer," Rock and Soil Mechanics, vol. 44, no. 5, pp. 1319-1329, 2023. doi: 10.16285/j.rsm.2022.0910

[19] R. Wang, H. Xu, Y. Liu, P. Jiang, A. Zhou, and Y. Lv, "Study on the shear strength characteristics and source mechanism of early-age shield synchronous grouting materials," Acta Geotechnica, vol. 18, no. 7, pp. 3573-3583, 2023. doi: 10.1007/s11440-023-01799-3

[20] S. Wang, Z. Lin, X. Peng, X. Wang, G. Tu, and Z. Song, "Research and evaluation on Water-dispersion resistance of synchronous grouting slurry in shield tunnel," Tunnelling and Underground Space Technology, vol. 129, pp. 104679, 2022. doi: 10.1016/j.tust.2022.104679

[21] Zhang, C. Zhu, M. Kuang, T. Xu, X. Wang, T. Feng, and J. Xu, "Research on the optimization of proportions for synchronous grouting material in composite muck during shield tunneling," Journal of Cleaner Production, vol. 484, pp. 144320, 2024. doi: 10.1016/j.jclepro.2024.144320

[22] H. Zhou, Y. Zhang, W. Zhu, Q. Zhong, and X. Huang, "Optimisation of Synchronous Grouting Mix Ratio for Shield Tunnels," Applied Sciences, vol. 14, no. 10, pp. 4098, 2024. doi: 10.3390/app14104098

[23] K. Elbaz, S.-L. Shen, A. Zhou, Z.-Y. Yin, and H.-M. Lyu, "Prediction of Disc Cutter Life During Shield Tunneling with AI via the Incorporation of a Genetic Algorithm into a GMDH-Type Neural Network," Engineering, vol. 7, no. 2, pp. 238-251, 2021. doi: 10.1016/j.eng.2020.02.016

[24] M. R. Moghaddasi, and M. Noorian-Bidgoli, "ICA-ANN, ANN and multiple regression models for prediction of surface settlement caused by tunneling," Tunnelling and Underground Space Technology, vol. 79, pp. 197-209, 2018. doi: 10.1016/j.tust.2018.04.016

[25] L. Sha, and Y. Yang, "ANN-based Structure Optimization with Fatigue Reliability Constrains," Applied Mechanics and Materials. pp. 3128-3131, 2012. doi: 10.4028/www.scientific.net/AMM.204-208.3128

[26] C. Bao, Y. Pu, and Y. Zhang, "Fractional-Order Deep Backpropagation Neural Network," Computational Intelligence and Neuroscience, vol. 2018, 2018. doi: 10.1155/2018/7361628

[27] K. AL-Bukhaiti, Y. Liu, S. Zhao, and H. Abas, "An application of BP neural network to the prediction of compressive strength in circular concrete columns confined with CFRP," KSCE Journal of Civil Engineering, vol. 27, no. 7, pp. 3006-3018, 2023. doi: 10.1007/s12205-023-1542-6

[28] G. Bai, and T. Xu, "Coal mine safety evaluation based on machine learning: a BP neural network model," Computational Intelligence and Neuroscience, vol. 2022, no. 1, pp. 5233845, 2022. doi: 10.1155/2022/5233845

[29] H. Chen, Z. Sun, Z. Zhong, and Y. Huang, "Fatigue factor assessment and life prediction of concrete

based on Bayesian regularized BP neural network," Materials, vol. 15, no. 13, pp. 4491, 2022. doi: 10.3390/ma15134491

[30] R. Chen, J. Song, M. Xu, X. Wang, Z. Yin, T. Liu, and N. Luo, "Prediction of the corrosion depth of oil well cement corroded by carbon dioxide using GA-BP neural network," Construction and Building Materials, vol. 394, pp. 132127, 2023. doi: 10.1016/j.conbuildmat.2023.132127

Dynamic BERT-Reinforcement Learning Model for Intent Recognition in Medical Dialogue Systems

Chunjun Cheng[1], Shui Cao[2 *], Guangyan Tang[3], Fang Ma[3], Di Cui[4]
[1]College of International Education, Jinzhou Medical University, Jinzhou 121000, China
[2]College of Medical Humanities, Jinzhou Medical University, Jinzhou 121000, China
[3]School of Computer Science, Jinzhou Normal College, Jinzhou 121000, China
[4]Department of Radiology, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121000, China
E-mail: ShuiCaoo@outlook.com
[*]Corresponding author

Keywords: medical dialogue system, BERT model, reinforcement learning, intent recognition, dynamic reasoning

Received: July 22, 2025

This study proposes a dynamic reasoning model of medical dialogue intentions that integrates BERT and reinforcement learning, aiming to solve the recognition difficulties caused by complex multiple rounds of interaction contexts and changeable user intentions in medical scenarios. Although the traditional BERT model is excellent in semantic modeling, it has limitations such as poor adaptability and static response strategy in the face of dynamic changes in intention expression in medical dialogue. Therefore, this paper introduces reinforcement learning mechanism, and realizes dynamic intention reasoning and policy optimization through state modeling, reward function and policy network. The experimental results highlight the robustness of our model in complex and dynamic medical dialogue scenarios. In high-complexity intent recognition tasks, our model achieved an accuracy improvement of 8.5%. Moreover, in extended multi-round dialogues, the BRL model demonstrated a significant increase in recognition accuracy—from 32% in the 70th round to 60% in the 140th round. This performance was notably better than that of the BLN model, which achieved about 40% accuracy. These improvements underscore the effectiveness of integrating reinforcement learning to adapt to evolving user intents and provide more accurate and contextually relevant responses in long-duration medical dialogues. In the sensitivity analysis of reward function, different reward functions have a significant impact on the model performance. Among them, RWA and RWF perform best when the weight numbers are 2 and 4, with an accuracy rate of more than 70%, while RWN and RWS are often below 40%. To sum up, the model combining BERT and reinforcement learning not only improves semantic understanding capabilities, but also realizes dynamic strategy adaptation, providing an efficient and intelligent intentional reasoning solution for medical dialogue systems.

Povzetek: Študija predlaga model, ki združi model BERT in okrepljeno učenje za boljše prepoznavanje namenov v večkrožnih medicinskih pogovorih, pri čemer se dinamično prilagaja spreminjajočim se uporabniškim namenom in opazno izboljša natančnost.

1 Introduction

With the continuous development of artificial intelligence technology, the application of Natural Language Processing (NLP) in the medical field has gradually deepened, and the medical dialogue system has become one of the research hotspots [1]. This system is designed to provide services such as diagnostic suggestions, symptom analysis and health consultation through natural language communication with patients. However, compared with general dialogue systems, medical dialogue has higher professionalism and accuracy requirements. It must accurately identify user intentions and adjust interaction strategies promptly to achieve efficient and reliable communication. Building a medical dialogue system with dynamic reasoning ability has become a key challenge in this context.

In recent years, the BERT (Bidirectional Encoder Representations from Transformers) model has been widely used in intent recognition tasks because of its excellent performance in semantic understanding [2]. BERT can effectively capture semantic context relationships through deep bidirectional language modeling, providing a powerful semantic representation basis for medical intention recognition [3]. However, BERT still has the problem of insufficient response when faced with real medical scenarios with frequent changes in dialogue context and changeable user intentions. Its static feature modeling mechanism makes

it difficult to cope with the policy adjustment requirements in dynamic interaction, which limits the model's performance in practical applications.

To solve the above problems, the introduction of the Reinforcement Learning (RL) mechanism has become an effective path [4]. Reinforcement learning can dynamically adjust the strategy according to the dialogue state and environmental feedback so that the system can continuously optimize the accuracy and decision-making efficiency of intentional reasoning in interaction with users [5]. By integrating reinforcement learning with the BERT model, the system can not only retain the advantages of BERT in semantic understanding but also enhance its reasoning ability in complex dialogue scenarios with the help of the adaptive characteristics of reinforcement learning and improve the overall intelligence level of the medical dialogue system.

Although the integration of BERT's powerful semantic understanding and adaptive reinforcement learning is indeed innovative, it is crucial to place our model in the context of current research. The most advanced existing models, such as BERT based intent recognition systems, perform well in static understanding of user input, but are difficult in terms of the dynamics of medical conversations, especially in multi round conversations. On the other hand, reinforcement learning (RL) models have achieved success in dynamic decision-making, but often lack the powerful semantic foundation required for accurate medical intent recognition. The method in this article combines the advantages of both, using BERT for in-depth semantic understanding and RL for dynamically adapting to the constantly changing nature of medical conversations. This hybrid approach not only improves the accuracy of intent recognition by enhancing the system's ability to understand context, but also optimizes decision-making strategies. Compared with traditional BERT based systems, the model proposed in this paper has achieved significant improvements in high complexity intent recognition, with a performance improvement of 8.5% in multi round medical conversations. Compared to models that only use reinforcement learning, our method benefits from a stronger semantic foundation, resulting in responses that are more context relevant and accurate.

This study aims to construct a dynamic reasoning model of medical dialogue intent that integrates BERT and reinforcement learning. The model can flexibly identify and reason user intent according to c, context changes in medical dialogue by introducing a state representation module, action decision mechanism, and reward feedback system. In this paper, systematic research will be carried out from the aspects of model structure design, strategy training methods, and experimental verification to improve the response intelligence and semantic accuracy of the medical dialogue system and provide theoretical support and a technical path for building a more humanized and efficient, intelligent medical service system.

2 Theoretical basis and related research

2.1 BERT and reinforcement learning algorithm theory

BERT is a pre-trained language model based on the Transformer structure, and its core advantage lies in the simultaneous capture of context information through a bidirectional encoder, thereby generating semantically rich word vector representations [6, 7]. The BERT pre-training process includes two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP) so that it can understand the deep semantic structure of the language [8]. As a basic model in natural language processing, BERT performs well in tasks such as question-answering systems, text classification, and named entity recognition. It is especially suitable for modeling complex semantic relationships and user intention recognition in medical dialogue systems.

However, BERT is essentially a static encoder model, and its inference mechanism relies on fixed parameters and offline training data, lacking responsiveness to user behavior changes in real-time interactions [9]. In medical dialogue scenarios, user intentions often change dynamically with the deepening of symptom descriptions or the adjustment of problem feedback, which puts forward higher adaptability requirements for the model. Therefore, relying solely on BERT to be competent for the intention recognition task with strong strategy and heavy context dependence in multi-round dialogue is difficult. To enhance the flexibility and adaptability of the system, it is necessary to introduce a dynamic mechanism that can handle sequence decision-making and environmental feedback to make up for the shortcomings of BERT in real-time interactive modeling [10].

As a learning mechanism centered on the interaction between agent and environment, RL is suitable for dealing with tasks with delayed feedback and state transition characteristics [11]. In the dialogue system, reinforcement learning can realize the joint optimization of intention recognition and dialogue strategy by constructing state space, action set, and reward function [12]. Specifically, the model can select the optimal response strategy according to the current user input (state) and continuously adjust the strategy parameters through user feedback, thereby forming the optimal intention reasoning path during training. The introduction of reinforcement learning improves the system's response to dynamic changes and enables the model to have the ability of online learning and strategy iteration.

Integrating BERT and reinforcement learning can achieve complementary semantic understanding and policy decision-making advantages. In the system architecture, BERT is a semantic encoder to provides semantic representation input for dialogue state modeling. At the same time, reinforcement learning guides the model to make optimal response judgments

in the policy selection module [13, 14]. This fusion method enhances the model's accuracy in medical semantic understanding. It optimizes the intention recognition strategy through continuous interaction, making it more aligned with the diversity and dynamics of patient expressions in real medical scenarios. Through this multi-level and collaborative-driven design idea, the intelligent reasoning ability of the medical dialogue system can be effectively improved, and more reliable and humanized support can be provided for intelligent health consultation.

2.2 Current status of healthcare conversational intent in BERT and reinforcement learning

Currently, medical dialogue systems are widely used in intelligent consultation, disease screening, and health consultation scenarios. One of their core tasks is to identify user intentions accurately. However, due to the complexity of technical terms and diverse expressions in the medical field, traditional intention recognition methods are often difficult to meet high precision requirements [15]. Many systems rely on rule templates or shallow classification models for intention recognition. Such methods have limited expressiveness in the face of semantic ambiguity or multiple rounds of dialogue. They are difficult to cope with patients' contextual changes and dynamic demands during the expression process, resulting in the system response lacking flexibility and semantic depth.

With the development of deep learning, pre-trained language models such as BERT have been introduced into medical intent recognition tasks, which greatly improves the system's ability to understand natural language semantics [16]. BERT is trained through many unsupervised corpora, has strong context modeling and semantic abstraction capabilities, and performs well in medical questions and answers, medical record summaries, and other tasks. When applied to medical conversations, BERT can effectively capture key information in patient statements, thus improving the accuracy of intent classification. However, existing studies mostly use BERT as a static feature extraction tool, ignoring the dynamic characteristics of user intention evolving within the context of medical dialogue. This often makes it difficult for the model to accurately track the transfer and development of user intention in actual interaction [17].

To make up for the shortcomings of static modeling, researchers gradually try to introduce reinforcement learning into medical dialogue systems and use its decision optimization ability to improve the interactive intelligence of the system. Reinforcement learning enables the system to learn when to confirm transfer or ask in-depth questions in multiple rounds of dialogue by constructing dialogue state space and reward mechanism and then dynamically adjusting the intention recognition strategy [18]. This method has shown positive effects in some medical scenarios, such as driving system strategy optimization through user feedback so that intention recognition relies on semantic features and considers interaction history and behavioral feedback. However, this kind of research is still relatively preliminary, and the generalization ability and training efficiency of reinforcement learning in high-dimensional medical semantic space still face challenges.

Therefore, combining the semantic understanding advantages of BERT with the strategy dynamic optimization ability of reinforcement learning is considered an important path to improving the performance of medical dialogue intention recognition [19]. By building a linkage mechanism, the system can not only make full use of BERT for fine semantic modeling but also realize continuous self-adjustment and optimization of intention recognition strategies with the help of reinforcement learning to be closer to complex interaction scenarios in actual diagnosis and treatment. The exploration of this direction at home and abroad is gradually deepening. Related research focuses on key issues such as model collaborative mechanism design, state representation selection, and multi-round dialogue task adaptation, which lays the foundation for building a medical dialogue system with adaptive reasoning ability

3 Establishment of dynamic reasoning model of medical dialogue intention based on BERT and reinforcement learning

3.1 Design and implementation of model framework

This study proposes a dynamic reasoning model of medical dialogue intention that combines BERT and reinforcement learning, aiming at improving the dialogue system's semantic understanding ability and dynamic reasoning ability in multiple rounds of interaction [20, 21]. The overall model architecture consists of two main modules: semantic understanding and state representation module, policy decision-making and intention reasoning module. The model design follows the three stages of "perception-decision-feedback". The system can use online reasoning and strategy self-optimization by combining the pre-trained language model BERT with the reinforcement learning strategy network while understanding user semantics [22]. This model is especially suitable for the interaction needs of complex user intentions, changeable semantic expressions, and long feedback chains in medical scenarios. The formula of the user input semantic vector extraction function is shown in (1).

$$h_t = BERT_\theta(u_t) \quad (1)$$

Where ht represents the semantics of the user input statement at time t, ut represents the user input text of the t round, BERTθ represents the pre-trained language model, and d represents the hidden dimension of the BERT output. The multi-round semantic state construction formula is shown in (2).

$$s_t = F_{state}(s_{t-1}, h_t) \quad (2)$$

Among them, st represents the state of the dialogue at time t, st−1 represents the state of the previous round, ht represents the semantic vector of the current round, and Fstate represents the state update function.

The reason for choosing this model is that traditional static intent recognition models, such as classifier or RNN-based structures, are difficult to adapt to the needs of dynamic intent evolution in medical dialogue. Although BERT has advantages in semantic modeling, it lacks the ability of strategic selection in dialogue. Reinforcement learning is good at optimal strategy learning in a dynamic environment [23]. Therefore, integrating the two can build an intelligent system that can understand semantics and dynamically reason, improving the accuracy and rationality of user experience and system response. In addition, the model also introduces the context state tracking mechanism to make the model have the ability of "memory" and enhance the semantic coherence modeling of multiple rounds of dialogue. The flow chart of designing a dynamic reasoning model of medical dialogue intention integrating BERT and reinforcement learning is shown in Figure 1.

The system begins by processing the original text of the medical dialogue, generating dynamic context vectors through data preprocessing, which includes tokenization and cleaning. These vectors are then input into the BERT model, which extracts high-dimensional semantic features that represent the underlying meaning of the user input. The semantic vectors produced by BERT are fed into the reinforcement learning (RL) module, where they are used to construct the current state representation. In this process, the RL module evaluates the state and makes action selections based on predefined policy networks. These actions drive the process of intent recognition by identifying the most relevant intents from the dialogue context, which are then used to refine the system's response strategy. The output from this interaction is a decision or response that is contextually informed by both semantic understanding and the RL-based dynamic reasoning process. The recognition results are evaluated by intent classification, and the reward signal is fed back to update the policy network, thus forming a closed-loop learning mechanism to optimize the accuracy of intent classification continuously. Finally, the system generates responses and outputs diagnoses or suggestions based on the classification results, realizing end-to-end intelligent medical dialogue intention recognition and response. This process realizes the deep integration of semantic understanding, policy optimization, and dynamic feedback and reflects the intelligent reasoning ability of the model.

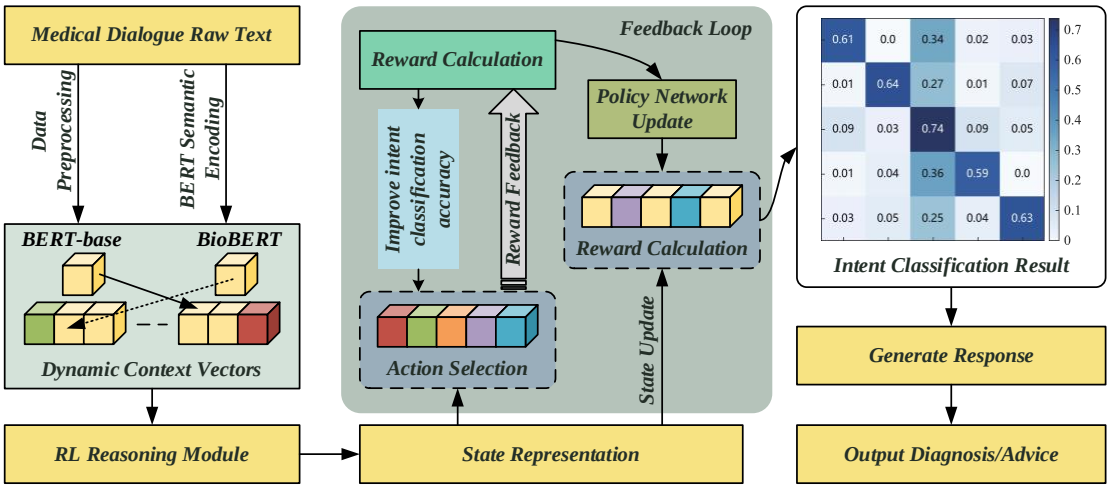


Figure 1: Design flow chart of medical dialogue intention dynamic reasoning model integrating BERT and reinforcement learning

This model is a fusion architecture that can combine semantic understanding and behavioral decision-making into a linkage system, avoiding the stage break in the "understanding-classification-response" process of the traditional model and continuously optimizing the reward mechanism in the interaction between the policy network and the environment, so that the model forms the optimal intention recognition path in the long-term interaction process [24]. This capability is particularly suitable for scenarios that include complex strategic operations such as confirmation, questioning, and guidance in medical conversations. The model's design supports soft state transition and explicit feedback

embedding, which improves the system's processing ability of unstructured medical expressions. The joint embedding representation function formula is shown in (3).

$$z_t = F_{joint}(h_t, s_{t-1}) \quad (3)$$

Among them, zt represents the semantics-policy at time t, ht represents the current round semantic vector extracted by BERT, st−1 represents the dialogue state of the previous round, and Fjoint represents the linkage function. The probabilistic state transition function formula is shown in (4).

$$s_t = (1-\lambda)\cdot s_{t-1} + \lambda \cdot z_t \quad (4)$$

Where st represents the state at the current time, λ represents the state fusion coefficient, st−1 represents the state at the previous time, and zt represents the current semantic-policy embedding. To verify the actual effect of the model, this paper constructs a simulated medical dialogue scenario including multiple rounds of question and answer, including the stages of user description of symptoms, system inquiry, user clarification, and systematic reasoning suggestions. When the user first describes "I've been dizzy recently", the system judges that the user may have neurological symptoms through state representation and then further refines the intention in multiple rounds of interaction, from "dizziness" to "whether it is accompanied by tinnitus" and "whether it lasts for a long time" etc. Sub-intention levels. Finally, the most likely dialogue path is judged through the strategy network, and diagnostic suggestions or medical advice are provided. The two modules included in the model-semantic understanding and state representation module, policy decision and intention reasoning module, work together in this process and constitute the key supporting structure of the complete reasoning process [25].

3.2 Semantic understanding and state representation module

The core of the semantic understanding and state representation module is the BERT model, which transforms users' natural language input into high-dimensional semantic vector representation as the input basis for subsequent decision modules [26]. In medical conversations, the language patients use is often ambiguous, non-normative, and highly context-dependent, so the context-aware ability of BERT is particularly important. We can obtain a more accurate and semantically consistent dialogue representation by splicing the patient input with the dialogue context and feeding it into the BERT encoder. In addition, the module also uses the representation of [CLS] bits as a global semantic feature, which plays a digest role in subsequent state modeling. The multi-round dialogue context stitching function formula is shown in (5).

$$T_t = Concat(u_{t-1}, u_t) \quad (5)$$

Where Tt represents the complete input text of the t round, ut−1 represents the dialogue history text, ut represents the current round user input, and Concat represents the text stitching operation. The output formula of intent classification is shown in (6).

$$y = softmax(Wv_T + b) \quad (6)$$

Where W denotes the weight matrix, b denotes the bias term, and y denotes the intent class probability distribution. In terms of state representation, this study designs a joint representation method, which integrates BERT semantic encoding of the current round of dialogue, state embedding of historical dialogue, and user feedback signal. This method not only retains the current input high-level semantic information but also abstractly models the dialogue history by memorizing the network structure to realize the modeling of contextual semantic continuity [27]. At the same time, a multi-layer perceptron is introduced to transform and dimension compression of the state vector to adapt to the state space requirements of the reinforcement learning strategy network. After this treatment, the system state has "current semantics" in the linguistic sense and "historical logic" in interactive behavior. The flow chart of dialogue state representation of semantic coding and state modeling is shown in Figure 2.

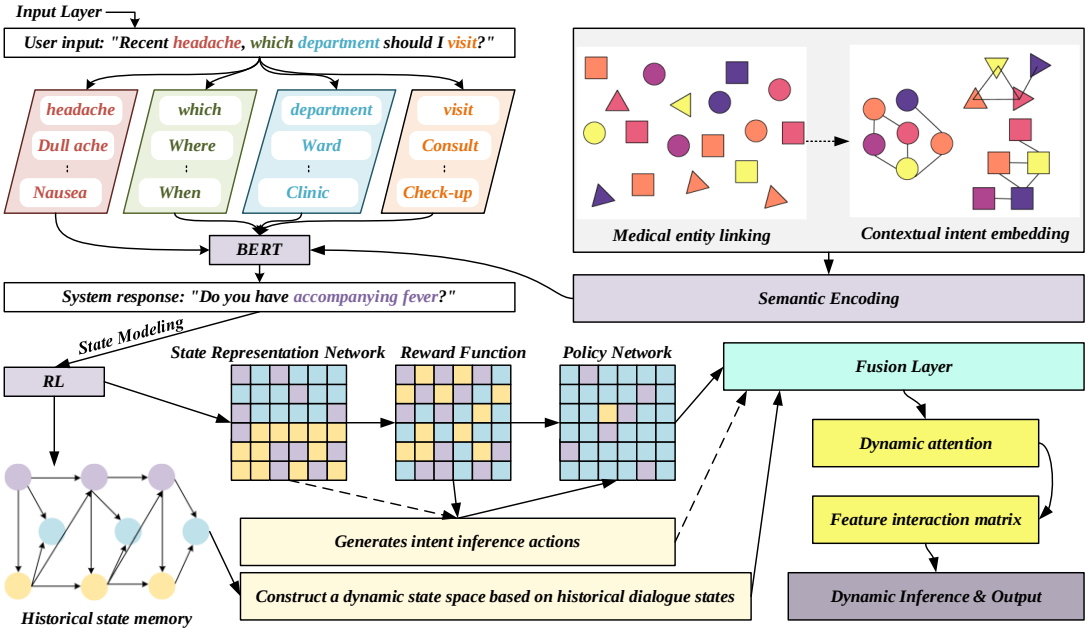


Figure 2: Flowchart of dialogue state representation integrating semantic coding and state modeling

The system first receives user input and embeds keywords semantically through BERT. Then, the system links medical entities and embeds contextual intentions in the semantic coding stage to enhance the understanding of users' semantics. Subsequently, the system uses the historical dialogue state to construct a dynamic state space. It generates intent reasoning actions through the reinforcement learning module's state representation network, reward function, and policy network. Finally, the fusion layer realizes the deep fusion of semantics and inference state through the dynamic attention mechanism and feature interaction matrix, outputs dynamic inference results, and generates responses, such as "Do you have accompanying fever?". The overall process implements a closed-loop dialogue intention recognition system from semantic understanding and state modeling to reasoning decision-making, improving the model's accuracy and context sensitivity in multiple rounds of medical dialogue.

We also introduce a dialogue context selection strategy based on the attention mechanism to improve the accuracy of representation and context understanding ability. This strategy dynamically determines which parts of the historical round are more critical to the current intention judgment, thus avoiding redundant information interfering with the model judgment. For example, the "history of allergies" mentioned by a user earlier is very critical under the intention of "rash", but it has less impact on the intention of "headache". By introducing a contextual attention mechanism, the model can automatically and selectively memorize or forget information, thus improving state vectors' discrimination and intention expression ability. The formula of the semantic coding sequence of the historical wheel is shown in (7).

$$L_{intent} = -\sum_{i=1}^{C} y_i log y'_i \quad (7)$$

Where yi denotes the true label, y'i denotes the prediction probability, and C denotes the total number of categories. The MDP five-tuple definition formula is shown in (8).

$$M = (S, A, P, R, \gamma) \quad (8)$$

Among them, S denotes the state space, A denotes the action space, P denotes the state transition probability, R denotes the reward function, and γ denotes the discount factor. The function of this module in the whole system is mainly to provide stable and informative semantic state input for policy decision-making. Through BERT's strong semantic coding of input information and historical state modeling, the model can generate a semantic state vector with accurate expression and a clear hierarchy for each round of dialogue. The design of this module ensures that the system can have a complete understanding of the user's semantic intention before the policy is executed, thus providing a solid foundation for downstream reasoning and response.

3.3 Policy decision and intention reasoning module

The strategy decision-making and intention reasoning module mainly comprises three parts: reinforcement learning strategy network, action set, and reward mechanism, which selects the optimal intention reasoning path in a dynamic environment. This module takes the semantic state vector as input. It outputs the

optimal action that should be taken, such as confirming the intent, requesting more information, or responding directly through the policy network [28]. The adopted strategy network is a deep Q network structure, which combines empirical playback and target network mechanism to improve training stability and strategy convergence efficiency. The model makes action decisions according to the current state in each round of dialogue to realize the dynamic adjustment of intention. The formula for calculating the DQN target value is shown in (9).

$$c_t = \alpha \cdot y_t + (1-\alpha) \cdot c_{t-1} \quad (9)$$

Where ct represents the cumulative confidence, α represents the update rate, and yi represents the current prediction probability ct−1 represents the confidence. The information gain is used for intent discrimination equation as shown in (10).

$$IG = H(Y) - H(Y \mid X) \quad (10)$$

Where H(Y) represents the label prior entropy, and H (Y | X) represents the conditional entropy. In action design, this module divides possible system behaviors into a variety of strategic actions, such as "confirming current intention", "asking clarification questions", "changing topic guidance," or "entering diagnosis mode". Each action closely corresponds to the actual medical consultation strategy, which makes the system more realistic and interactive in the reasoning process [29]. This design effectively supports the hierarchical evolution process of complex intentions, such as developing from "pain" to "abdominal pain" and then specifically to "severe pain in the right lower abdomen" and judging it as "appendicitis risk", and guiding the dialogue to a meaningful direction through the strategy network. Evolution. The intention recognition target aggregation reward function formula is shown in (11).

$$R_t = \sum_{j=1}^{k} \gamma^j \quad (11)$$

Where Rt denotes the aggregate reward of the current dialog round, γ denotes the discount factor, and k denotes the prediction round number window. The design of the reward mechanism is one of the key parts of this module. This paper sets the real-time reward value for each action, and the global task completion reward and error penalty are introduced. For example, if the user confirms that the system infers the correct intent, a high reward is given; If the system deviates from the user's intention or causes the user's disgust, a penalty is imposed. In addition, the model also calculates the total return through evaluation indicators such as accuracy and satisfaction when the dialogue is completed, driving strategy learning to move closer to better goals. This reinforcement mechanism makes the model pursue short-term response correctness and optimize long-term interaction paths [30].

During the implementation process, the user described "the abdomen is a little uncomfortable recently", and the system initially predicted the intention of "stomach disease". However, according to the historical dialogue and state representation, it was found that the user had previously mentioned "lower abdomen", and the strategy network guided the system to ask further "whether the pain is located in the right lower abdomen", and then accurately reasoned it as possible appendicitis and recommended medical treatment. This multi-round, state-driven strategy reasoning ability is the advantage of this module. With the introduction of this module, the whole model can continuously interact with users, dynamically update intention judgment, and optimize dialogue strategy.

4 Experimental results and analysis

This study used the MedDialog Chinese Medical Dialogue dataset as the experimental data source, covering multiple rounds of consultation dialogues in different medical departments in China, suitable for intent recognition and dialogue strategy modeling tasks. Before model training, the dataset undergoes several preprocessing steps, including text cleaning, tokenization, and filtering out irrelevant conversations. Then divide each conversation into separate rounds to capture the dynamic nature of the conversation and ensure that the model can understand the constantly changing user intentions. In addition, to address potential data imbalance issues, especially for rare medical intentions, we use data augmentation techniques to generate synthetic dialogues, ensuring strong training and coverage for frequent and infrequent intention categories. After standardization, the data are divided into training sets, verification sets, and test sets to ensure the accuracy and reliability of experimental evaluation. The experiment was conducted on a deep learning server running the Ubuntu system, with the hardware configuration consisting of an NVIDIA RTX 3090 graphics card, an Intel Xeon processor, and 128GB of memory. The software environment was built using Python, leveraging the PyTorch framework. The BERT model was loaded using the transformers library, and reinforcement learning was implemented via OpenAI Gym, which provided a stable environment for training and evaluating the model. This setup ensures robust performance during the training process, with an emphasis on the integration of BERT and reinforcement learning for medical dialogue intent recognition. Model performance comparisons are shown in Table 1.

Table 1: Model performance comparison

Models	Accuracy (%)	Recall rate (%)	F1 value (%)	Inference time (ms)
BiLSTM	82.1	80.5	81.3	35
BERT	89.3	88.9	89.1	58
BERT + RL	91.7	90.8	91.2	61
Our model	94.5	93.8	94.1	63

The model proposed in this paper leads all over the four indicators. The accuracy rate reaches 94.5%, 12.4% higher than the benchmark model and 5.2% higher than the standard BERT model. The increase in F1 value is 12.8% and 5%, significantly better than traditional and deep language models without integrated reinforcement learning. The inference time is slightly increased (5ms more than BERT), but the performance improvement far exceeds the cost increase, indicating that the introduced dynamic intent inference mechanism is extremely cost-effective. Especially in medical scenarios, accuracy precedes response speed, and model selection strategies are more reasonable.

This paper analyzes the influence of different model structures on the accuracy of intent recognition in order to evaluate this influence, and the results are shown in Figure 3.

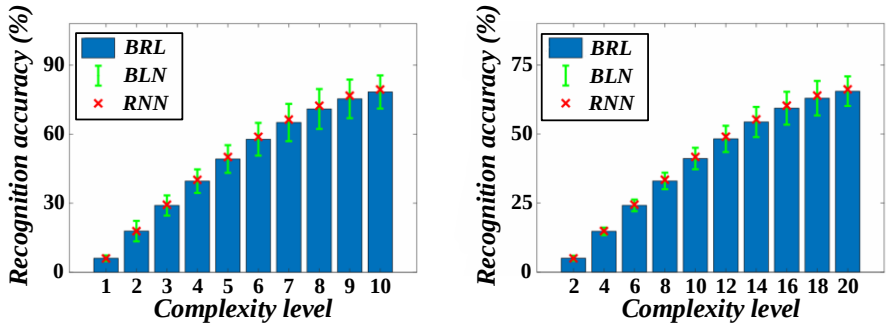


Figure 3: Influence of model structure on intention recognition accuracy

As you can see from the chart, in the range of complexity levels 1 to 10 in the left figure, the accuracy rate of the BRL model (fusing BERT and reinforcement learning) has increased from about 10% to nearly 90%, which is always significantly better than BLN (BERT only) and RNN, especially when the complexity is greater than 5, the gap widens. In the figure on the right, in the range of complexity levels 2 to 20, the recognition accuracy of the BRL model steadily rises to more than 75%, while BLN and RNN finally stay at about 65% and 60%. The overall display shows that the BRL model has stronger expression and reasoning ability when dealing with high-complexity intentions, and the structural optimization significantly improves the system's understanding accuracy of complex semantics.

This paper analyzes the changes in intention recognition accuracy under different conversation rounds to verify whether the model can maintain the stability and accuracy of intention understanding in multiple rounds of medical answering and answering. The results are shown in Figure 4.

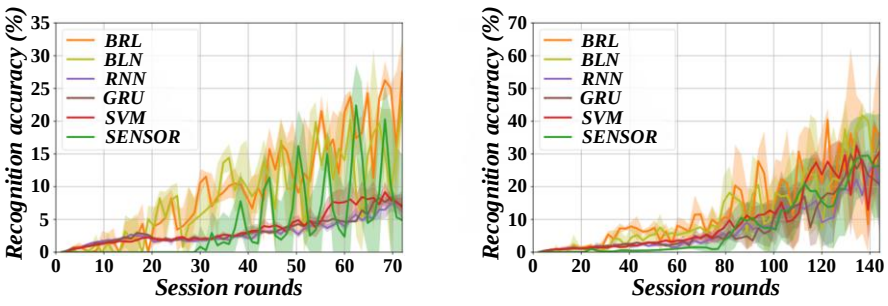


Figure 4: Changes of intent recognition accuracy under different session rounds

According to the data in the figure, the BRL in the left figure achieves a recognition accuracy of about 32% in 70 rounds of dialogue, which is significantly higher than that of BLN and other traditional models such as RNN and GRU. In the figure on the right, as the number of session rounds is expanded to 140 rounds, the recognition accuracy of BRL is further improved to about 60%. In contrast, the highest accuracy of BLN,

SENSOR, and SVM remains at about 40%. This shows that BRL has a stronger intention understanding and adaptability in long-round interaction scenarios, and the reinforcement learning mechanism significantly enhances the model's contextual reasoning ability and robustness.

This paper analyzes the sensitivity of reinforcement learning reward function weights on model performance to explore the influence of three weight adjustments of "semantic matching," "context consistency," and "action selection confidence" in different reward functions on model performance and analyze the model's dependence on the policy feedback structure. The results are shown in Figure 5.

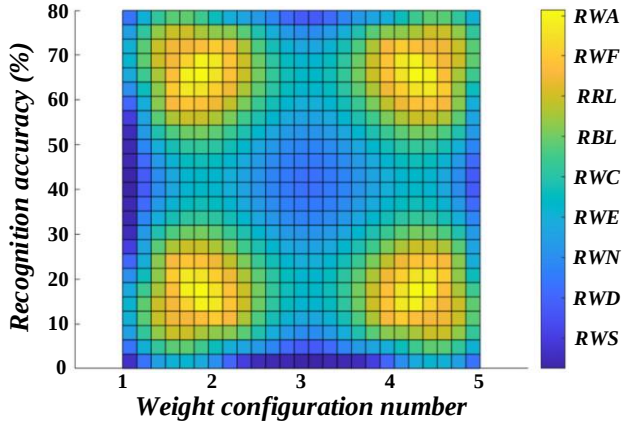


Figure 5: Sensitivity analysis of reinforcement learning reward function weights to model performance

It can be seen from the figure that there are obvious differences in the influence of different reward functions on the accuracy of model intent recognition under different weight configuration numbers. RWA and RWF perform best when the weight configuration numbers are 2 and 4, and the recognition accuracy rate reaches more than 70%, which is close to the yellow area in the figure. However, the accuracy of RWN and RWS models is concentrated between 20% and 40% in most configurations, and the performance is poor. Reward function design and weight allocation are highly sensitive to model performance. Reasonable selection of reinforcement learning reward mechanism is a key factor in improving model accuracy.

Table 2: Recognition accuracy under different intention complexity

Intent complexity level	Number of samples	BERT Accuracy (%)	Model accuracy in this paper (%)	Improvement (%)
Low	240	92.5	95.2	+2.7
Medium	310	88.1	91.6	+3.5
High	180	79.4	87.9	+8.5

The recognition accuracy under different intention complexity is shown in Table 2. The model shows significant advantages in high-complexity intent recognition. Especially in the "merge intent" scenario, the model's accuracy in this paper has increased by 8.5 percentage points, indicating that the reinforcement learning mechanism effectively enhances the multi-intent recognition ability of the model. The improvement in medium complexity intentions also reached 3.5%, reflecting that the dynamic reasoning strategy helps understand the semantic dependencies implicit in the context. Overall, the more complex the model is, the more obvious the advantages of this method are, and it is suitable for actual medical dialogue scenarios with multiple rounds of interaction and multi-intent recognition.

This article analyzes the relationship between the proportion of different intent categories in the data set and the classification accuracy of the model to evaluate the impact of data imbalance on model performance. The result is shown in Figure 6.

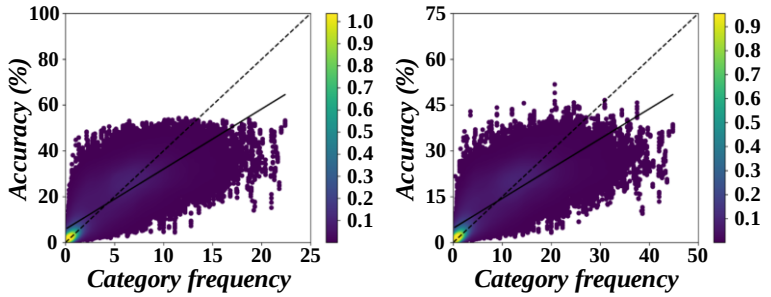


Figure 6: Relationship between intent category distribution and classification accuracy

In the figure on the left, when the category frequency is less than 5, most classification accuracy rates are concentrated between 0% and 40%. In contrast, when the frequency is increased to more than 15, the accuracy rate can reach more than 60% and even close to 90% in some areas. The figure on the right shows a similar trend. Still, the overall accuracy range is lower, mostly between 15% and 45%, indicating that the model's accuracy is significantly limited when dealing with low-frequency categories. This shows that the model has a better learning effect on high-frequency categories under the long-tail intent distribution. However, there are still obvious shortcomings in recognizing low-frequency categories, emphasizing the necessity of introducing mechanisms such as reinforcement learning to enhance intention recognition with few samples.

This paper compares the visualization of the model reasoning path before and after the dynamic reasoning module is enabled to compare the model's path changes during the reasoning process and show the dynamic reasoning module's optimization effect on the model decision path. The results are shown in Figure 7.

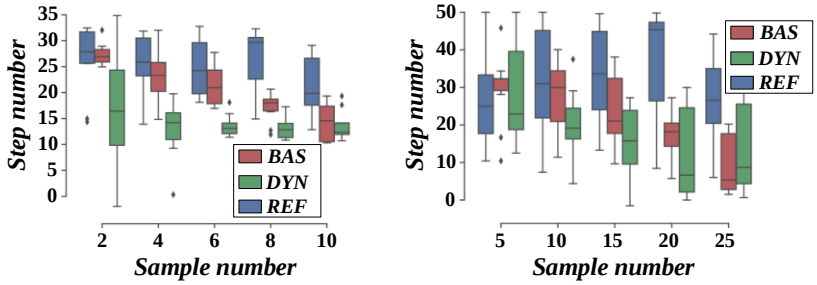


Figure 7: Visual comparison of model inference paths before and after the dynamic inference module is enabled

It can be observed in the figure that when the dynamic reasoning module is not enabled, the number of reasoning steps of the basic model (BAS) is generally high, and the reasoning path of multiple samples exceeds 30 steps. After the dynamic inference module is enabled, the inference steps of the DYN model in most samples are significantly shortened, and multiple samples are concentrated between 10 and 20 steps, which is significantly close to the reference path (REF) labeled by experts. For example, when the sample numbers are 2 and 10, the inference steps of the DYN model are about 15 and 10, respectively, while that of the BAS model is about 30 and 28, respectively. Overall, DYN is superior to BAS in inference efficiency and path compactness, indicating that the dynamic inference module can effectively optimize the decision path of the model and make it closer to the expert level.

To observe the changing trend of the loss value of the verification set with the rounds during the training process and judge whether the model is overfitted or under fitted, this paper analyzes the changing trend of the loss function of the model training rounds and the verification set, and the results are shown in Figure 8.

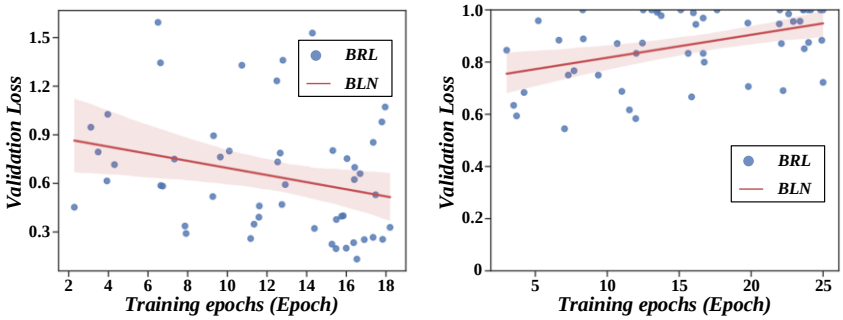


Figure 8: Model training rounds and verification set loss function change trend diagram

According to the data in the figure, the model (BRL) that combines BERT and reinforcement learning shows an overall downward trend in the verification set loss during the training process, from about 0.9 to a minimum of about 0.3. Although the fluctuation is large, the final convergence effect is obvious. Although the model using only BERT (BLN) has a slight downward trend in the left figure, the overall change is relatively gentle. In the right figure, there is even a phenomenon that the verification loss increases with the training rounds, from 0.8 to close to 1.0, showing the potential overfitting risk. This comparison result shows that the BRL model is more stable in the training process and has stronger generalization ability on the validation set. The reinforcement learning module has a significant optimization effect on the model training effect.

The trends of training rounds and model performance are shown in Table 3. From the perspective of the training process, the model's performance has improved significantly after 10 rounds. The peak accuracy rate reached 94.5% in the 15th round, and the loss value dropped to 0.21, which has tended to be stable. Although the loss declined in the 20th round, the accuracy rate fluctuated slightly, showing an overfitting trend. It shows that the model structure design is reasonable, the training efficiency is high, the optimal performance can be achieved within 15 rounds, and the training cost is well controlled. Using an early stop strategy in actual deployment is recommended to avoid overfitting.

This paper analyzes the changes in intent recognition accuracy in multiple rounds of user interactions to demonstrate the accuracy of intent recognition in each of the 10 rounds of user interactions and evaluate the stability and robustness of the system in long-term dialogues. The results are shown in Figure 9.

Table 3: Training rounds and model performance change trends

Training rounds	Accuracy (%)	Loss	Learning rate	Rate of convergence
5	85.6	0.42	1e-4	Non-convergent
10	91.2	0.29	1e-4	Accelerated descent phase
15	94.5	0.21	1e-4	Achieve a stable optimum
20	94.4	0.20	1e-4	Slight signs of overfitting

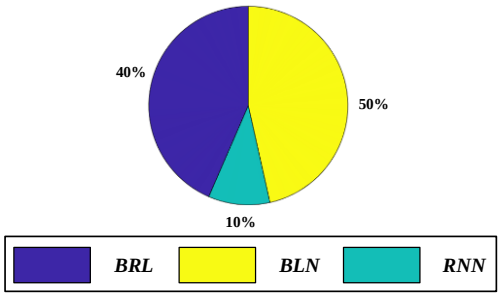


Figure 9: Change of intention recognition accuracy in multiple rounds of user interaction

As can be seen from the figure, the model (BRL) that combines BERT and reinforcement learning accounts for 40% of the accuracy improvement contribution in intention recognition in multiple rounds of interaction and performs better than other models. The BLN (using only BERT) model has the highest contribution of 50%. Still, its promotion is mainly concentrated in the first few rounds, and the overall stability is slightly inferior to BRL's. However, the contribution of the traditional RNN model is only 10%, indicating that its intention recognition ability in a multi-round interactive environment is significantly

insufficient. This result confirms that the dynamic reasoning mechanism proposed in this paper can more effectively improve the understanding and response accuracy of the model in complex interaction scenarios.

5 Conclusion

This study aims to enhance intention recognition in medical interactions by introducing a reinforcement learning strategy network and semantic state modeling, showing strong performance in experiments. However, it lacks real-world validation, and clinical testing with expert benchmarking is needed to address complexities like patient behavior variability and real-time feedback. Challenges remain in handling rare medical conditions and ambiguous patient inputs due to limited training data. Future improvements will focus on data augmentation and integrating domain-specific datasets to overcome these issues. The specific conclusions are as follows:

(1) The experimental results show that the model proposed in this paper is better than traditional models such as BERT and BiLSTM in terms of accuracy, recall rate, and F1 value, with an accuracy rate of 94.5%, which is 12.4 percentage points higher than BiLSTM and 5.2 percentage points higher than BERT; The F1 value also increased from 89.1% of BERT to 94.1%, an increase of 5 percentage points. This result fully proves that the fusion of BERT and reinforcement learning strategy has an important effect on complex medical semantic modeling and intent reasoning.

(2) In multiple rounds of conversation testing, the BRL model reached an accuracy of 32% in the 70th round and increased to 60% in the 140th round, far exceeding the BLN model. In addition, in scenarios with high intention complexity, this model's accuracy rate reaches 87.9%, which is 8.5 percentage points higher than BERT. This shows that the model has good long-term interactive context tracking ability and complex intention understanding ability.

(3) In the sensitivity analysis of different reward functions, the accuracy rate of RWA and RWF strategies exceeds 70% when the weight configuration numbers are 2 and 4. In contrast, the RWN strategy without a reinforcement module is only about 30%-40%, showing that the reward mechanism profoundly impacts the quality of model decision-making. At the same time, the model has reached an accuracy rate of 94.5% in the 15th round of training, and the Loss has dropped to 0.21. The training convergence speed is fast, and the performance is stable. However, the BLN model was overfitted during the training process, and the verification loss increased from 0.8 to nearly 1.0, further highlighting the enhancement effect of reinforcement learning on generalization ability.

The dynamic reasoning model of medical dialogue intention combines BERT and reinforcement learning and has significant advantages in dealing with complex, long-round, and highly semantic-dependent medical dialogue tasks. The model not only realizes the linkage optimization of semantic understanding and reasoning strategies but also shows high accuracy, high robustness, and high training efficiency in actual tests, which can provide strong support for the implementation of intelligent medical service systems. In the future, research can further explore the directions of multi-modal input and multi-intention fusion decision-making to promote the development of a medical dialogue system to a more intelligent and humanized.

Funding

This work was supported by the Social Science Planning Fund Project of Liaoning Province (L24CYY013).

References

[1] Das, S., & Ambika, M. "Revolutionizing Indian Farming: Machine Learning-Powered NLP for Optimal Crop Recommendations," Procedia Computer Science, vol. 258, pp. 2040–2049, 2025, DOI: 10.1016/j.procs.2025.04.454.

[2] Abdullah, A. S., Geetha, S., Govindarajan, Y., Pranav, A. G. V., & Vinod, A. A. "Enhanced Information Retrieval Using Hybrid p-Norm Extended Boolean Models with BERT," Procedia Computer Science, vol. 258, pp. 3052–3061, 2025, DOI: 10.1016/j.procs.2025.04.563.

[3] Baruah, P., Dutta, B., Sarma, S. K., & Talukdar, K. "Named Entity Recognition in Assamese Language using two separate models: BiLSTM and BERT," Procedia Computer Science, vol. 258, pp. 242–251, 2025, DOI: 10.1016/j.procs.2025.04.262.

[4] Fu, Y., Han, J., Xu, Y., Liu, K., Lu, J., He, S., & Bo, X. "RL-GCL: Reinforcement Learning-Guided Contrastive Learning for molecular property prediction," Information Fusion, vol. 122, pp. 103208, 2025, DOI: 10.1016/j.inffus.2025.103208.

[5] Jafar, M. T., Yang, L.-X., & Li, G. "An innovative practical roadmap for optimal control strategies in malware propagation through the integration of RL with MPC," Computers & Security, vol. 148, pp. 104186, 2025, DOI: 10.1016/j.cose.2024.104186.

[6] Beniwal, R., & Saraswat, P. "A hybrid BERT-CPSO model for multi-class depression detection using pure hindi and hinglish multimodal data on social media," Computers and Electrical Engineering, vol. 120, pp. 109786, 2024, DOI: 10.1016/j.compeleceng.2024.109786.

[7] Chi, D., Huang, T., Jia, Z., & Zhang, S. "Research on sentiment analysis of hotel review text based on BERT-TCN-BiLSTM-attention model," Array, vol. 25, pp. 100378, 2025, DOI: 10.1016/j.array.2025.100378.

[8] Clavié, B., Cooper, N., & Warner, B. "It is all in the [MASK]: Simple instruction-tuning enables BERT-like masked language models as generative classifiers," Natural Language Processing Journal, vol. 11, pp. 100150, 2025, DOI: 10.1016/j.nlp.2025.100150.

[9] Gui, J., Zhou, Y., Yu, K., & Wu, X. "PSC-BERT: A spam identification and classification algorithm via prompt learning and spell check," Knowledge-Based Systems, vol. 301, pp. 112266, 2024, DOI: 10.1016/j.knosys.2024.112266.

[10] Guo, Y., Xie, Z., Chen, X., Chen, H., Wang, L., Du, H., Wei, S., Zhao, Y., Li, Q., & Wu, G. "ESIE-BERT: Enriching sub-words information explicitly with BERT for intent classification and slot filling," Neurocomputing, vol. 591, pp. 127725, 2024, DOI: 10.1016/j.neucom.2024.127725.

[11] Gupta, B. B., Gaurav, A., Arya, V., Attar, R. W., Bansal, S., Alhomoud, A., & Chui, K. T. "Advanced BERT and CNN-Based Computational Model for Phishing Detection in Enterprise Systems," CMES - Computer Modeling in Engineering and Sciences, vol. 141, no. 3, pp. 2165–2183, 2024, DOI: 10.32604/cmes.2024.056473.

[12] Haurogné, J., Basheer, N., & Islam, S. "Vulnerability detection using BERT based LLM model with transparency obligation practice towards trustworthy AI," Machine Learning with Applications, vol. 18, pp. 100598, 2024, DOI: 10.1016/j.mlwa.2024.100598.

[13] He, B., Zhao, R., & Tang, D. "CABiLSTM-BERT: Aspect-based sentiment analysis model based on deep implicit feature extraction," Knowledge-Based Systems, vol. 309, pp. 112782, 2025, DOI: 10.1016/j.knosys.2024.112782.

[14] He, G., Lin, C., Ren, J., & Duan, P. "Predicting the emergence of disruptive technologies by comparing with references via soft prompt-aware shared BERT," Journal of Informetrics, vol. 18, no. 4, pp. 101596, 2024, DOI: 10.1016/j.joi.2024.101596.

[15] Hussain, A., Saadia, A., & Alserhani, F. M. "Ransomware detection and family classification using fine-tuned BERT and RoBERTa models," Egyptian Informatics Journal, vol. 30, pp. 100645, 2025, DOI: 10.1016/j.eij.2025.100645.

[16] Jamshidi, S., Mohammadi, M., Bagheri, S., Najafabadi, H. E., Rezvanian, A., Gheisari, M., Ghaderzadeh, M., Shahabi, A. S., & Wu, Z. "Effective text classification using BERT, MTM LSTM, and DT," Data & Knowledge Engineering, vol. 151, pp. 102306, 2024, DOI: 10.1016/j.datak.2024.102306.

[17] Liao, W., Liu, Z., Dai, H., Wu, Z., Zhang, Y., Huang, X., Chen, Y., Jiang, X., Liu, D., Zhu, D., Li, S., Liu, W., Liu, T., Li, Q., Cai, H., & Li, X. "Mask-guided BERT for few-shot text classification," Neurocomputing, vol. 610, pp. 128576, 2024., DOI: 10.1016/j.neucom.2024.128576.

[18] Mao, X., Li, Z., Li, Q., & Zhang, S. "BERT-DXLMA: Enhanced representation learning and generalization model for english text classification," Neurocomputing, vol. 622, pp. 129325, 2025, DOI: 10.1016/j.neucom.2024.129325.

[19] Mirtaheri, S. L., Pugliese, A., Movahed, N., & Shahbazian, R. "A comparative analysis on using GPT and BERT for automated vulnerability scoring," Intelligent Systems with Applications, vol. 26, pp. 200515, 2025, DOI: 10.1016/j.iswa.2025.200515.

[20] Onan, A., & Alhumyani, H. A. "FuzzyTP-BERT: Enhancing extractive text summarization with fuzzy topic modeling and transformer networks," Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 6, pp. 102080, 2024, DOI: 10.1016/j.jksuci.2024.102080.

[21] Panoutsopoulos, H., Espejo-Garcia, B., Raaijmakers, S., Wang, X., Fountas, S., & Brewster, C. "Investigating the effect of different fine-tuning configuration scenarios on agricultural term extraction using BERT," Computers and Electronics in Agriculture, vol. 225, pp. 109268, 2024, DOI: 10.1016/j.compag.2024.109268.

[22] Qin, S., & Zhang, M. "Boosting generalization of fine-tuning BERT for fake news detection," Information Processing & Management, vol. 61no.4, pp. 103745, 2024, DOI: 10.1016/j.ipm.2024.103745.

[23] Rahman, M. H., Uddin, M. A., Ria, Z. F., & Rahman, R. M. "Optimizing BERT for Bengali Emotion Classification: Evaluating Knowledge Distillation, Pruning, and Quantization," CMES - Computer Modeling in Engineering and Sciences, vol. 142no.2, pp. 1637–1666, 2025, DOI: 10.32604/cmes.2024.058329.

[24] Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. "Optimizing Airline Review Sentiment Analysis: A Comparative Analysis of LLaMA and BERT Models through Fine-Tuning and Few-Shot Learning," Computers, Materials and Continua, vol. 82no.2, pp. 2769–2792, 2025, DOI: 10.32604/cmc.2025.059567.

[25] Tawil, A. A., Almazaydeh, L., Qawasmeh, D., Qawasmeh, B., Alshinwan, M., & Elleithy, K. "Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT," Computers, Materials and Continua, vol. 81no.2, pp. 3395–3412, 2024, DOI: 10.32604/cmc.2024.057279.

[26] Teng, X., Zhang, L., Gao, P., Yu, C., & Sun, S. "BERT-Driven stock price trend prediction utilizing tokenized stock data and multi-step optimization approach," Applied Soft Computing, vol. 170, pp. 112627, 2025, DOI: 10.1016/j.asoc.2024.112627.

[27] Wan, B., Wu, P., Yeo, C. K., & Li, G. "Emotion-cognitive reasoning integrated BERT for sentiment analysis of online public opinions on emergencies," Information Processing & Management, vol. 61no.2, pp. 103609, 2024, DOI: 10.1016/j.ipm.2023.103609.

[28] Wang, H., Song, K., Jiang, X., & He, Z. "ragBERT: Relationship-aligned and grammar-wise BERT model for image captioning," Image and Vision Computing, vol. 148, pp. 105105, 2024, DOI: 10.1016/j.imavis.2024.105105.

[29] Wu, D., Yang, J., & Wang, K. "Exploring the reversal curse and other deductive logical reasoning in BERT and GPT-based large language models," Patterns, vol. 5no.9, pp. 101030, 2024, DOI: 10.1016/j.patter.2024.101030.

[30] Xu, L., & Wang, W. "Aspect-based sentiment classification with BERT and AI feedback," Natural Language Processing Journal, vol. 10, pp. 100136, 2025, DOI: 10.1016/j.nlp.2025.100136.

Dynamic Elderly Care Resource Allocation using SHBM-Tuned Deep Q-Networks (SHBM-RDQN)

Hua Jiang
School of social work, China Civil Affairs University, Beijing, 102600, China
E-mail: jianghua_jh630@outlook.com

Keywords: elderly care services, resource allocation, staff optimization, health management systems, dynamic honeybees mating-tuned resource based Deep Q-Network (DHBM-RDQN)

Received: August 29, 2025

The growing demand for elderly care services necessitates intelligent, efficient, and adaptive resource management strategies. This research presents a dynamic allocation strategy for optimizing elderly care resources using reinforcement learning (RL), addressing the complexity of nursing home resource management. This research proposes a Dynamic Honeybees Mating–tuned Resource-based Deep Q-Network (DHBM-RDQN) for elderly care resource allocation. The approach models the care environment as a Markov Decision Process (MDP), where states capture patient acuity levels, staff availability, daily admissions, and care workload, while actions correspond to dynamic allocation of nursing staff and support resources. The system uses a comprehensive Elderly Care Staffing & Quality Dataset, comprising time-series records at daily or shift-level granularity from multiple long-term care facilities. Preprocessing includes Z-score normalization and missing value imputation, while feature extraction applies Independent Component Analysis (ICA) and Discrete Wavelet Transform (DWT) to capture latent health patterns and temporal fluctuations. The DHBM-RDQN employs a Deep Q-Network with two fully connected layers (256 and 128 neurons) and ReLU activations, trained using the Adam optimizer. The Honeybee Mating optimization layer dynamically tunes learning rate, exploration parameters, and reward weights to prevent premature convergence. Experimental evaluation were implemented in python. Results show 96.5% accuracy, 90.5% resource efficiency, 1.0 s response time, and an adaptability score of 0.925, demonstrating robust adaptability under fluctuating patient demand and staffing. This research introduces a novel RL framework combining deep learning and bio-inspired optimization to achieve superior performance, rapid decision-making, and improved care quality in dynamic long-term elderly care environments.

Povzetek: Raziskava predstavlja dinamično strategijo razporejanja virov v domovih za starejše z okrepljevalnim učenjem (RL) in bio-navdihnjeno optimizacijo, ki izboljša učinkovitost, prilagodljivost ter kakovost oskrbe ob nihanjih povpraševanja in razpoložljivosti osebja.

1 Introduction

One of the major problems facing the world today is the aging population particularly in the developing countries whereby by 2050, the numbers of individuals aged 65 years and above are projected to have doubled. This demographic transition goes a long way in creating a high demand of efficient healthcare systems and facilities among the aged. To have long-term sustainability of the workforce and quality care delivery, effective nursing staffing and resources are required. Introducing data-driven solutions and real-time analysis in the context of health management may be used to resolve these issues and guarantee efficiency of resources and quality of the services [1, 2]. Besides addressing complex medical, emotional, and social needs, nursing homes and assisted living facilities are currently challenged by the persistent issues, including the rising operational costs, the shortage of staff members, and the growing need to provide personalized care [3]. These difficulties emphasize the necessity of clever, responsive solutions that would help balance the quality of care and optimal resources allocation [4].

The use of AI technology, including ML, NLP, and intelligent robots, by more healthcare organizations to improve QOC services and medical resource utilization is growing. Such advantages enable regular chances of transformation within a knowledge-oriented healthcare setting [5]. Information technology offers a channel through which the AI healthcare service resources can interact stakeholders be it patients, clinicians, pharmaceutical and insurance companies, and the hospital serves as the central point [6]. The field of healthcare has now been transformed by AI through such applications as DL models and rule-based systems. Under AI is RL, which has drawn increased attention to its application in medical applications. Better personalized medicine, better use of resources in hospitals, more optimizations in therapy, and accuracy of diagnostics can be developed with the help of AI. AI

algorithms have made it faster to discover diseases, enhanced the pace of accuracy during diagnostics, and offered useful information to make decisions. Moreover, automation, which has been facilitated by artificial intelligence, has enhanced administrative procedures and reduced practitioner workload that had enhanced efficiency in operations [7].

The latest studies have given emphasis on the benefits of RL to enhance operation in healthcare. RL-based frameworks have so far been implemented to control hospital bed capacity, enhance surgical schedule, and enable emergency response planning. These applications were more flexible when compared to traditional optimization models [8]. Dynamic allocation models based on ML was found to enhance patient safety, emergency response and overall quality of service in the context of elderly care. These models are based on round the clock monitoring of the daily health indicators to determine the needs of care and staff [9]. Healthcare planners are usually presented with multidimensional, complex issues that present scheduling issues, resource allocation, and strategic decision-making. These problems will necessitate smart and flexible systems that can manage complexity and uncertainty in the operating environments [10].

1.1 Aim of the research and objective

This research aims to develop an intelligent and adaptive framework for optimizing elderly care resource allocation in nursing homes. The objective is to dynamically manage nursing staff and support resources using a reinforcement learning-based model, DHBM-RDQN, which integrates Deep Q-Networks with Honeybee Mating optimization. By modeling the environment as a Markov Decision Process and leveraging feature extraction techniques like ICA and DWT, the framework seeks to improve accuracy, resource efficiency, response time, and adaptability, ensuring robust and efficient care under fluctuating patient demand and staffing conditions.

1.2 Key Contribution of the research

- To propose a dynamic RL method DHBM-RDQN to optimize elderly care service resources, improving staff allocation efficiency and service quality.
- To utilize a comprehensive elderly care staffing & quality dataset, consisting of time-series records on patient demand, staffing, operational constraints, and care quality outcomes from multiple long-term care facilities.
- To utilize preprocessing to standardize heterogeneous elderly care data and extract meaningful patterns from multi-source variables, improving the quality of inputs for reinforcement learning.
- To formulate the elderly care environment as an MDP, capturing patient acuity, staff availability, and care demands for structured state-action-reward learning.
- To demonstrate that the proposed method outperforms traditional rule-based approaches, providing more adaptive, reliable, and effective resource allocation in elderly care settings.

2 Related works

The new ERAS in the fog environment specially developed to be used in healthcare was introduced [11]. To achieve effective resource management, ERAS applied prediction algorithms and current allocation of resources. Unlike the old algorithms, ERAS optimizes ARU and Level of Load Balancing and gives the minimal Makespan. The two primary contributions were the use of optimized RL for load balancing and resource allocation in the fog environment, as well as the optimization of RL hyperparameters by PSO.

ElHealth, an IoT-focused approach that tracks hospital room usage by patients and modifies medical staff accordingly, was presented [12]. It forecasted the amount of space needed and suggested ways to assign experts by data prediction algorithms. Additionally, ElHealth offered proactive human resources elastic speedup and multi-level predictive elasticity of human resources to control resource utilization. Using data, the model was tested and demonstrated encouraging outcomes, lowering waiting times by as much as 96.71%.

System performance and efficiency were improved by streamlining the HRM process and lowering effort [13]. The HRM system model was built on a BPNN driven by DL algorithms. The model provided the best optimization effect over traditional models and converged the quickest. According to the results of the LOO approach, the model converges at 60 epochs and was 2.76% more accurate than other models, with an accuracy of 88.72%.

MILP that used the Gurobi optimization solver to reduce healthcare costs and enhance patient care was presented [14]. Due to the implementation of two different scenarios, two distinct optimal solutions were obtained. The first produced an ideal solution with a 0.0% solution gap and an objective value of 844.0. With an objective score of 539.0, the second solution validated the model's dependability. There were no discernible discrepancies between the two approaches' best-bound scores, suggesting optimal solutions with reasonable tolerances.

The application of system dynamics modeling was looked to improve quality of service in the system of healthcare [15]. Errors were reduced and work pressure was stabilized, but when the patient arrival rates were lowered, it went against accessibility objectives. Errors were decreased and service capacity was increased by increasing human resources, especially experienced staff. Long-term increases in service quality required striking a balance between system costs and staff efficiencies. Patient satisfaction and operational performance could both be improved by incorporating dynamic modeling into management procedures.

Two important issues in healthcare team-based planning of resources were addressed by ML and stochastic optimization [16]. To satisfy demands and cut expenses, it focused on forecasting patient workloads and allocating healthcare team resources as efficiently as possible. Before allocating patients to available teams and balancing workloads in narrow decision-making, the model employed a deep multi-task learning technique to forecast workloads for various patient categories. Multi task learning performed better than traditional prediction techniques and takes stochastic variables and randomness into account for better access to healthcare.

The suitability of NCD models in hospital wards and the efficiency of ML techniques in improving nurse staffing were assessed [17]. The highest accuracy in predicting the suitability of NCD systems and the adequacy of nurse staffing in 39 inpatient wards were shown by RF algorithm. Functional nursing and complete patient care models were the most commonly utilized care delivery models in wards.

A scheduling system of nurse schedules was developed automatically using open-source operational research techniques [18]. All hard criteria and the majority of the soft constraints were satisfied by the system, which produced schedules in less than a minute. Nurses and management staff could communicate in real time by the computer-generated schedules, which were more flexible and optimally accustomed. The system was applied in many wards and was cost-effective, efficient, and easy to use. It was also updated frequently with new policies and nurse staff.

Meanwhile, the application of predictive modeling and workforce analytics based on the AI was provided as an innovative strategy distribution of human resources in the healthcare industry [19]. It included data on clinical outcomes, measure of operational efficiency, patient acuity, and past staffing trends. Having the best chance to forecast staffing requirements in various clinical contexts, the technique increased the use of resources, patient satisfaction, and operational cost-effectiveness. These workforce management systems are AI-based, which creates a balance between the quality of care and the effectiveness of the organization.

In addressing the high-dimensional constraints of conventional SARSA, an approximation sate-action-reward-state-action (ASARSA) algorithm has been proposed to optimize resource allocation in energy harvesting (EH)-multiple-input multiple-output (MIMO) communication systems [20]. Comparing experimental results to SARSA and Q-Learning, they demonstrate reduced error, faster convergence, and increased throughput. All things considered, ASARSA shows improved scalability, precision, and efficiency for practical EH-MIMO applications.

Queue assessment model was developed for assessing walk-in outpatients in a public hospital that did not have appointment scheduling [21]. The model evaluated wait times over a seven-week period using DEA. Doctor/personnel and consultation time were among the outputs; the latter was the non-discretionary output. Excel VBA programming was used to provide the dynamic framework for ongoing queue monitoring.

The scheduling procedure for ED patients was optimized with the use of deep RL [22]. DQN was the foundation of the algorithm, which was intended to reduce waiting times and penalties for patients who were emergencies. In terms of reducing waiting time and penalties, the research demonstrated that RL performs better than dispatching rules. Deep RL successfully applied in ED applications, as this research has shown, especially when it comes to supporting decision-makers in the dynamic setting of an ED. Table 1 shows the summary table for related works in healthcare resource management.

Table 1: Summary of related works in healthcare resource management

Research No.	Technology Used	Domain	Dataset	Performance Metrics
[11]	ERAS, RL, PSO	Fog-based healthcare resource management	Real-time healthcare resources in fog environment	ARU, Load Balancing Level, Makespan
[12]	ElHealth, IoT, Prediction algorithms	Hospital room usage and staffing	Patient occupancy & staff allocation	Waiting time reduction (up to 96.71%)
[13]	BPNN, Deep Learning	HRM process optimization	Healthcare staffing and task data	Accuracy 88.72%, convergence at 60 epochs
[14]	MILP, Gurobi optimizer	Cost reduction & patient care	Healthcare operational scenarios	Solution gap 0%, Objective values 844 & 539
[15]	System Dynamics Modeling	Quality of service improvement	Healthcare operational & staffing data	Error reduction, stabilized work pressure, service capacity
[16]	ML, Stochastic optimization, Deep multi-task learning	Team-based healthcare resource planning	Patient workloads & healthcare team data	Forecasting accuracy, workload balancing efficiency
[17]	RF, ML	Nurse staffing and NCD models	39 inpatient wards	Prediction accuracy for NCD suitability and staffing adequacy
[18]	Open-source Operational Research Techniques	Nurse scheduling	Nurse shift & scheduling constraints	Feasibility of schedules, satisfaction of hard/soft constraints, efficiency

[19]	AI, Predictive modeling, Workforce analytics	Strategic HR allocation	Clinical results, operational efficiency, staffing trends	Staffing prediction accuracy, resource utilization, patient satisfaction
[20]	SERVQUAL model	Patient service satisfaction	22 service quality variables	Statistical correlation between service dimensions and patient satisfaction
[21]	Queue assessment model, DEA, Excel VBA	Outpatient waiting time	Walk-in patient data over 7 weeks	Wait times, doctor/personnel consultation time
[22]	Deep RL, DQN	ED patient scheduling	Emergency department patient arrivals & tasks	Reduced waiting times, penalties for emergency patients

2.1 Research gap

Despite the continual advancements in healthcare resource utilization, several gaps in HRM, care coordination, and service delivery still remain. The existing approaches of ERAS [11] and ElHealth [12], for example, are primarily focused on predictive allocation and elasticities, which tend not to adequately integrate multiple sources of time sensitive patient and operational data in a holistic adaptive decision-making framework. Traditional optimization frameworks such as BPNN-based HRM [13] and MILP models [14] provide improvements to efficiency but rely on static historical datasets that constrain their ability to adapt to real-time dynamic care settings. Similarly, system dynamics models for HRM and ML-driven workload forecasts [15, 16] provide useful means of improving planning but do not consolidate dynamic patient health variation with dynamic nurse staffing availability within an integrated framework. Approaches addressing nurse rostering and HR [17, 18, and 19] focus on scheduling and operational efficiency but have not utilized RL as a means of continuously optimizing resource allocation policies under uncertainty and risk. Assessments of service quality and access / ED patient scheduling using deep RL [20] [21] [22], while demonstrating promise, are also limited in scope and do not offer extrapolation opportunities for multi-institution elder care contexts. This highlights the need for a dynamic, RL-based approach DHBM-RDQN that integrates heterogeneous, temporal, and operational data for real-time elderly care resource optimization.

3 Methodology

The proposed methodology for dynamic allocation of elderly care resources was detailed in this section. It covers dataset characteristics, preprocessing steps including Z-score normalization and missing value imputation, feature extraction via ICA and DWT, and MDP formulation for RL-based decision making. The hybrid DHBM-RDQN algorithm combines resource based deep Q-learning (RDQN) and Dynamic Honeybee Mating optimization (DHBM), which can be used to find efficient, adaptive, and robust resource allocation given real-time changes in patient demand and staff availability. The general way of flow of the methodology is provided in Figure 1.

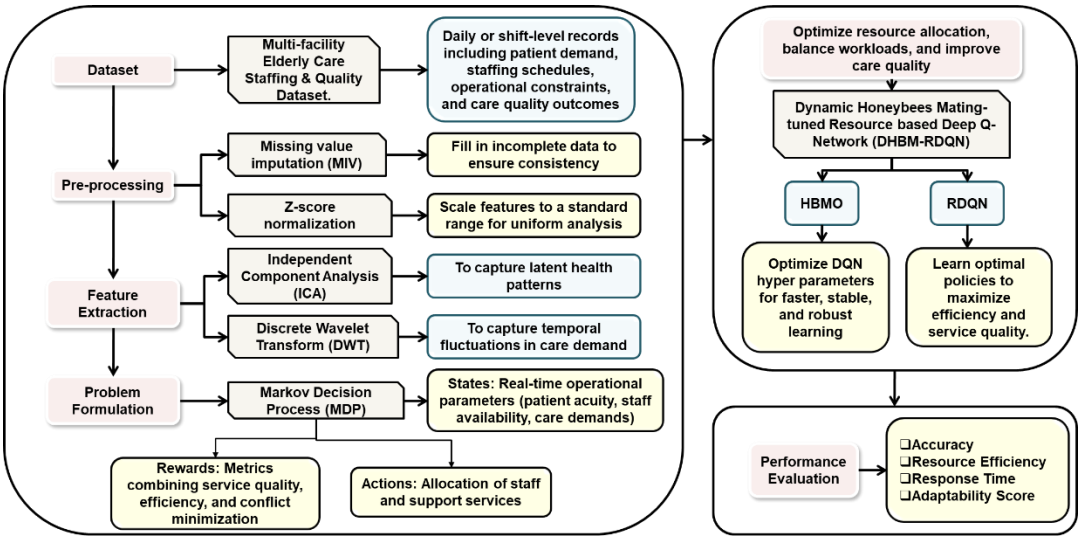


Figure 1: Methodology flow for RL based dynamic allocation of elderly care resources

3.1 Dataset

The elderly care staffing & quality dataset from Kaggle (https://www.kaggle.com/datasets/zara2099/elderly-care-staffing-and-quality-dataset/data) comprises time-series data collected from multiple long-term care facilities, capturing comprehensive information on staffing, workload, operational constraints, and care quality outcomes. The data is also organized on a daily or shift level, which allows examining the dynamic resource distribution and efficiency of operations. It contains both continuous and discrete attributes, which capture the patient demand, nurse supply, time factors, cost of operations, and quality indicators, which makes it appropriate in RL-based optimization of resources in the elderly care setting. The most important characteristics of the dataset are provided in Table 2.

Table 2: Key features of elderly care dataset

Category	Feature Examples
Demand / Workload	Daily resident census, admissions, discharges, transfers, care minutes index, patient acuity scores, unplanned events (falls, rapid response calls, wound care)
Supply / Staffing	Staff headcount by role: Registered Nurse (RN) , Licensed Practical Nurse (LPN), CAN (Certified Nursing Assistant) and skill mix percentages, overtime/agency/float pool hours, staff absences, planned leave
Temporal & Context	Shift type, day of week, month, season, holiday indicators, pay cycle flags, outbreak flags, extreme weather conditions
Operational Constraints & Costs	Staffing ratios, maximum hours per staff, hourly wages, overtime/agency premiums, budget caps
Quality & Outcomes	Pressure-injury cases, medication errors, readmissions, complaints, satisfaction scores

3.2 Preprocessing of Elderly care data for RL

Preprocessing is a critical step in preparing the elderly care dataset for RL, ensuring that raw, heterogeneous data is transformed into a consistent and reliable format suitable for modeling. By applying the below two techniques, the dataset becomes standardized, complete, and free from distortions. This preprocessing step ultimately enhances the stability, accuracy, and generalizability of the proposed RL framework for dynamic elderly care resource allocation.

1) Z-score Normalization: Z-score normalization for preprocessing of heterogeneous variables such as patient health indicators, staffing schedules, and emergency response times was utilized to achieve uniform scaling. The goal of Z-score normalization is to put the variables onto a standardized scale to better stabilize the reinforcement learning algorithm during convergence. Unlike Min–Max scaling, this technique centers data on zero and scales it based on Standard deviation. This means that when learning a policy, attributes like patient BP, which span a higher numeric range, will not be relied upon more than attributes like staff-to-patient ratios, which span a smaller range. The Z-score for a given variable x_i is calculated using Equation (1).

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

Where x_i represents the original values, μ is the mean of the variable, σ denotes Standard deviation. Following normalization, all variables are on a common scale with a mean of zero and unit variance, easing the comparison of features in the state-action-reward space of the MDP.

2) Missing Value Imputation (MVI): In real world elderly care data, missing data points are unavoidable due to missing staff logs, unrecorded evaluation of patient health signs, or delayed documentation of emergency situations. If not handled appropriately, these untreated missing values can introduce bias and diminish the robustness of RL outcomes. To mitigate this issue, the research decorates and applies MVI methods that are relevant to the data type, for example, a mean substitution for continuous variables (i.e., response time, physiological measures, HR, BP etc.,), median substitution for skewed measures, and mode imputation for the variable types of staff roles, shift types, and service ratings. This method would help to substitute the missing information with statistically representative values and maintain the variations in the data distribution to not bias the patterns that are needed for effective policy discovery. By systematically imputing missing values, the dataset remains complete and coherent.

The dataset contained missing values ranging from 2.8% to 11.4% across different features, with staffing logs and emergency records having the highest gaps. Missing values in time-series sequences were handled using forward-fill and backward-fill interpolation to maintain temporal continuity and avoid abrupt gaps in data flow, ensuring realistic sequential patterns for RL training. Additionally, a bias evaluation was performed by comparing statistical properties (mean, variance, and range) before and after imputation, and the results showed no significant deviation, confirming that the imputation strategy did not distort the dataset.

Algorithm 1: Data preprocessing

```
# Step 1: Load Dataset
dataset = load_csv("elderly_care_dataset.csv")
# Step 2: Z-score Normalization
for feature in continuous_features:      mean_val = mean(dataset[feature])
   std_val = std(dataset[feature])
   dataset[feature] = (dataset[feature] - mean_val) / std_val

# Step 3: Missing Value Imputation (MVI)
for feature in dataset.columns:
   missing_percentage = calculate_missing_percentage(dataset[feature])

   if missing_percentage > 0:
      if feature in continuous_features:
         dataset[feature].fillna(mean(dataset[feature]), inplace=True)
      elif feature in skewed_features:
dataset[feature].fillna(median(dataset[feature]), inplace=True)
      elif feature in categorical_features:
         dataset[feature].fillna(mode(dataset[feature]), inplace=True)

# Step 4: Time-series Continuity Handling
for feature in time_series_features:
   dataset[feature] = forward_fill(dataset[feature])      dataset[feature] = backward_fill(dataset[feature])
# Step 5: Bias Evaluation
for feature in dataset.columns:
   compare_statistics(original_data[feature], dataset[feature])
# Step 6: Output Preprocessed Dataset
save_csv(dataset,
"preprocessed_elderly_care_dataset.csv")
```

3.3 Feature extraction

Elderly care data are important per se since they unveil concealed correlations, and temporal dynamics, which can be used in the reinforcement learning process. ICA isolates independent latent signals among several correlated variables of health and workload, whereas Discrete Wavelet Transform (DWT) detects multi-resolution temporal variations which enhance state-space descriptions and facilitate the learning of policies around the dynamic allocation of resources.

ICA was chosen to separate overlapping signals in elderly care data into independent components, capturing hidden trends better than variance-based methods like PCA. DWT was employed to handle non-stationary temporal patterns, providing both time and frequency information to detect short-term anomalies and long-term workload trends. Together, ICA and DWT generate a rich feature set that enhances the model's dynamic decision-making capabilities.

1) Independent Component Analysis (ICA): ICA was used to find some latent independent factors from the multidimensional care of the elderly dataset, in particular when strong associations exist among the variables. For instance, the health indicators of the patient, such as HR, BP, and level of mobility, exhibit strong correlations and interdependencies, masking the independent signals of health deterioration or health stabilization. These techniques decompose the observed dataset X into strains of statistically independent components in accordance with Equation (2).

$$X = AS \quad (2)$$

Where X is the observed data matrix, A is the mixing matrix, and S is the independent components matrix in Equation (2). This technique aims to approximate both A and S such that the extracted components in S are maximizing independence. In this research, the independent components are interpreted as latent health risk indicators, workload dynamics, and hidden care demand patterns to add to the state-space representation of the reinforcement learning model. By having the independent signals, the proposed model would be able to learn more optimal policies for dynamic resource allocation while not creating redundancies associated with raw correlated features. The first 5 independent components capturing the most significant latent health and workload patterns were retained post-ICA for the state-space representation in the RL model.

2) Discrete Wavelet Transform (DWT): WT is yet another feature extraction method like Fourier method, but unlike Fourier method, it is able to provide time while also keeping frequency data. WT allows one to capture temporal and frequency-specific change in the elderly care data set, beyond what has already been shown. WT is especially useful for time series data that is irregular and non-stationary such as the emergence of

patterns for staff, health fluctuations for patients, and overall workload patterns for staff.

The DWT (Equation (3)) decomposes a time-series signal f(t)into a set of approximations and details using scaled and shifted versions of a mother wavelet function ψ, t is a continuous time variable over which the signal is measured.

$$\text{DWT}(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right) dt \quad (3)$$

Where the scale and translation parameters are a and b , respectively, which regulate frequency resolution and temporal localization.

The Daubechies (db4) wavelet was used for the Discrete Wavelet Transform (DWT) to extract multi-resolution temporal features from the elderly care dataset. The wavelet decomposition was used for identifying multi-resolution features, including energy, entropy, and dominant frequency bands for time-series emergency logs, patient monitoring, and daily care demand variation logs to detect short-run anomalies (such as spikes in the number of emergencies) and long-run care demand trends (such as seasonality in health needs of patients).

3.4 Markov Decision Process (MDP)

To optimize the allocation of elderly care service resources, the environment is modeled as an MDP. By offering a mathematical foundation for sequential decision-making, MDP enables the RL agent to discover the best course of action for allocating resources in real time, contingent on the condition of the care facility. An MDP is defined by the tuple (S, A, R, P) where S represents the state, A represents the action, R is the reward, and P denotes the transition probability.

At each time step t, the system examines the current state s_t, selects an action $b_t \in A$, and receives a reward r_t as the environment transitions to a new state s_{t+1}. Although exact transition probabilities may not be fully known, the RL framework can still learn effective policies through interactions with the environment.

State Formulation: The state captures real-time operational parameters of the elderly care facility, including patient acuity levels, staff availability, and care demands. Let $B\wp_{ft}^{BV}$ be the total number of patients with the acuity level $bv = \{1, \dots, BV\}$ in $\wp_{ft}$, and let $\wp_{ft}$ be a collection of patients waiting to have a procedure on medical resource type f at time t. This will characterize the condition at time t . f can perform a variety of treatments, as previously indicated. Next, in $\wp_{ft}$, let RT_{ft}^{tt} be the total number of patients in $\wp_{ft}$with treatment type $tt = 1, \dots, TT$ where TT is the total number of treatment types, that need to be processed on f. As a result, the following Equation (4) represents the status $S_t^f$of each resource group g at t.

$$S_t^f = \left\{\left(\frac{B\wp_{ft}^1}{|\wp_{ft}|}, \dots, \frac{B\wp_{ft}^{BV}}{|\wp_{ft}|}\right), \left(\frac{RT_{ft}^1}{|\wp_{ft}|}, \dots, \frac{RT_{ft}^{TT}}{|\wp_{ft}|}\right)\right\} \quad (4)$$

Standardizing each element to a ratio between 0 and 1 ensures that variables with different scales do not disproportionately influence the learning process. Lastly, Equation (5) displays s_t.

$$s_t = \{S_t^1, \dots, S_t^f\} \quad (5)$$

State Vector Example:

For a facility with 3 resource types$(f = 1,2,3)$, 2 acuity levels $(BV = 2)$, and 2 treatment types $(TT = 2)$, the state vector at time t can be represented as:

$$s_t = \{S_t^1, S_t^2, S_t^3\}$$
$$= \{(0.4,0.6), (0.3,0.7), (0.5,0.5), (0.2,0.8), (0.6,0.4), (0.3,0.7)\}$$

Here, each tuple in S_t^f shows the normalized ratio of patients by acuity and treatment type for each resource

Action formulation: Actions represent the dynamic allocation of elderly care resources, such as assigning a staff member to a patient or scheduling a care task. Equation (6) describes the existence of $\wp_{ft}$ and $b_t$and also denotes the selected patient or task allocation for each resource type *f* at time t. The RL agent selects these actions to optimize the overall service efficiency.

$$b_t = \{\wp_{1t}, \dots, \wp_{ft}\} \quad (6)$$

Number of Actions per Time Step: The total number of actions depends on the number of patients and resource types. If $\wp_{ft}$ contains N patients for resource f, and there are F resources, then at each time step:

$$\text{Total actions} = \sum_{f=1}^{F} |\wp_{ft}|$$

Each action represents assigning a specific patient to a resource for a treatment step.

Reward formulation: The reward function is designed to incentivize efficient resource consumption and minimize patient waiting times. Specifically, the reward r_t at time t is given as the negative weighted sum of waiting times for the assigned patients (Equation 7).

$$r_t = -1 \times c_j\left\{\sum_{k=1}^{k-1} wt_{jk}\right\} \quad (7)$$

Where wt_{jk} is the waiting time of patient j for treatment step k and c_j is the corresponding weight reflecting priority or care severity. Negative rewards (-1) ensure that the RL agent is encouraged to reduce waiting times while balancing staff allocation and care quality.

3.5 Dynamic Honeybees Mating-tuned Resource based Deep Q-Network (DHBM-RDQN)

Combining the RDQN with DHBM establishes a hybrid method for learning and optimization that simultaneously leverages policy learning with an adaptive search. While RDQN uses an approximation of the Q-function for efficient resource allocation in high-dimensional elderly care contexts, DHBM increases focused exploration by introducing mutation-based diversity to mitigate premature convergence. Taken together, RDQN with DHBM allows the agent to learn stable yet flexible policies that can respond optimally and rapidly to changing patient needs and workforce availability and maintain optimal scheduling performance.

3.5.1 Resource based Deep Q-Network (RDQN)

The proposed approach employs an RDQN to dynamically optimize elderly care resource allocation. In traditional Q-Learning, the agent seeks to learn Q-function, $Q(s_t, b_t)$, which symbolizes the probable cumulative reward for action taking b_t in state s_t and following the correct policy thereafter, receiving the reward r_t. The optimal policy selects actions that maximize future returns. The Q-function is iteratively updated by Equation 8.

$$Q(s_t, b_t) \leftarrow (1-\alpha)Q(s_t, b_t) + \alpha[r_t + \gamma Q(s_{t+1}, b_{max})] \quad (8)$$

Where the learning rate is represented by α. The discount factor γ was accounting for the uncertainty of future rewards. In standard Q-Learning, $Q(s_t, b_t)$is stored in a table, which becomes infeasible for state-action spaces of high-dimensional such as those in elderly care facilities.

To address this, RDQN replaces the Q-table with a DNN, $Q_\theta(s, b)$, parameterized by θ, which approximates Q-values for all state-action pairs. During training, the network predicts $Q_\theta(s_t, b_t)$ in a forward pass, while the resulting experience tuple (s_t, b_t, r_t, s_{t+1}) is stored in a replay buffer. The loss function is then minimized to update the network settings (Equation (9)).

As an estimate of future reward, the goal is to utilize the observed reward, r_t, to get an improved approximation of the Q-value. The discount factor γ takes into consideration the uncertainty of future rewards, whereas α represents the pace of learning. In DQN, a DNN $Q_\theta(s, b)$ is used in place of the Q-table, and back-propagation is used to train the parameters θ. An RL problem does not provide a ground truth, in contrast to supervised learning. The network's current estimate, $Q_\theta(s_t, b_t)$, for the Q-value is determined in a forward pass. Executing the corresponding action yields the tuple (s_t, b_t, r_t, s_{t+1}) that is then saved in an experience buffer. However, the network parameters θ are adjusted using a cost function rather than individual Q-values.

$$J_\theta = \frac{1}{2}\left(Q_\theta(s_t, b_t) - \left(r_t + \gamma \max_{b_j} Q_\theta(s_{t+1}, b_{max})\right)\right)^2 \quad (9)$$

In Equation (9), a ground truth γ is approximated using the definition $r_t + \gamma \max_{b_j} Q_\theta(s_{t+1}, b_{max})$, and the network is trained using the resulting cost J_θ. However, training can be unstable because the same network is used to estimate both the Q-value and its target. To mitigate instability, the RDQN incorporates mechanisms such as experience replay and target network updates, ensuring stable convergence. By integrating RDQN with the state representations derived from ICA and Wavelet features, the agent learns to allocate staff and care resources efficiently, minimizing waiting times and maximizing service quality in elderly care facilities.

3.5.2 Dynamic Honeybees Mating (DHBM)

The DHBM algorithm is inspired by the natural mating behavior of a hive queen bee, which ensures exploration and exploitation in a solution space.

The DHBM mechanism in this framework functions as an adaptive optimizer that dynamically enhances the Q-learning process rather than operating independently. Instead of relying solely on epsilon-greedy or fixed exploration parameters, DHBM generates diverse policy candidates through mutation and mating strategies and evaluates them based on their fitness (resource allocation efficiency and waiting-time minimization). The best-performing candidate policies are then used to update the RDQN parameters, adjust exploration–exploitation balance, and tune hyperparameters such as learning rate and discount factor during training. This integration ensures that the Q-network avoids premature convergence, maintains policy diversity, and continuously adapts to shifting patient loads and staffing fluctuations in real time.

DHBM operates directly on the policy network, where it refines network weights during training. The experience replay buffer remains unchanged, and DHBM does not modify stored transitions. Instead, its role is to enhance policy optimization by improving convergence and decision accuracy.

In the context of this research, each bee represents a potential scheduling and resource allocation policy for elderly care facilities, where the fitness of each bee corresponds to its effectiveness in reducing patient waiting time and balancing caregiver workload. In the classic HBMO method, the best bee is designated a queen bee; the rest are regarded as drones. The mating with the queen occurs only probabilistically, defined in Equation (10).

$$\text{Prob}(C) = \exp(-\Delta(e)/T(l)) \quad (10)$$

Where $\text{Prob}(C)$denotes the probability of a drone contributing its genetic material (solution) to the queen's sperm theca, $\Delta(e)$ represents the absolute difference between the drone and queen, and $T(l)$ represents the queen's speed at iteration l. A higher probability of mating occurs when the queen is moving quickly or when the drone's fitness is similar to the queen's. After each iteration, the queen's speed decreases according to Equation (11). $T(l+1)$ in Equation (11) represents the queen bee's "speed" at the next iteration $l+1$.

$$T(l+1) = \alpha \times T(l) \quad (11)$$

Where $\alpha \in [0, 1]$is the speed decay factor. To prevent premature convergence and better adapt to the dynamic and uncertain care demands in elderly facilities, the DHBM introduces mutation strategies that diversify the search space. Specifically, at each iteration four distinct mutant vectors are generated using different mutation rules (Equations 12–16). This enhances global exploration and avoids being trapped in local optima when patient demand fluctuates unpredictably. To uniformly cover the whole searching region, the algorithm mutates vectors in each step by choosing four vectors ($W_{q1}, W_{q2,} W_{q3,} W_{q4,}$) from the original population as $W_{q1} \neq W_{q2} \neq W_{q3} \neq W_{q4}$.

$$W^{j}_{mutant1} = W^{j}_{q1} + E_1 \times \left(W^{j}_{queen} - W^{j}_{r3}\right) + E_1 \times (W^{j}_{q3} - W^{j}_{q4}) \quad (12)$$

$$W^{j}_{mutant2} = W^{j}_{queen} + E_2 \times \left(W^{j}_{q1} - W^{j}_{q2}\right) \quad (13)$$

$$W^{j}_{mutant3} = W^{j}_{q1} + E_3 \times \left(W^{j}_{q2} - W^{j}_{q3}\right) + E_3 \times (W^{j}_{q1} - W^{j}_{q4}) \quad (14)$$

$$W^{j}_{mutant4} = (W^{j}_{q1} + W^{j}_{q2} + W^{j}_{q3})/3 + \left(\beta_2 - \beta_1\right)\left(W^{j}_{q1} - W^{j}_{q2}\right) + (\beta_{3-}\beta_2)(W^{j}_{q2} - W^{j}_{q3}) + (\beta_1 - \beta_3)(W^{j}_{q3} - W^{j}_{q1}) \quad (15)$$

$$\beta_1 = \frac{\left|e(W^{j}_{q1})\right|}{\beta^*}, \beta_2 = \frac{\left|e(W^{j}_{q2})\right|}{\beta^*}, \beta_3 = \frac{\left|e(W^{j}_{q3})\right|}{\beta^*} \quad (16)$$

The function to be optimized is denoted by $\beta^* = |e\,(W^{j}_{q1}) + |e\left(W^{j}_{q2}\right) + |e\,(W^{j}_{q3})|$ and $e(W)$, the coefficients between 0 and 1 are denoted by E_1 to E_3, the mutant vector of the jth iteration associated with the ith mutant rule is denoted by $W^{j}_{mutant\ i}$, and the queen vector (which yields the best result) is denoted by W^{j}_{queen} at iteration j. In this research, the fitness function evaluates scheduling policies based on two objectives: minimizing weighted patient waiting time, and balancing workload distribution across caregivers. $e(W^{j}_{q1})$ is the magnitude of that fitness (always non-negative).

The position of the vector is changed to the upper and lower bounds of the control vector if any component of any mutant vector violates its constraint. Next, all of the mutant vectors' fitness functions are calculated and arranged using the descending technique. The mutation vector chosen as $W^{j}_{mut,Best}$ is the one with the lowest fitness function. In the following generation, the trial vector takes the place of the target vector if the cost of the mutant vector is lower than that of the target (Equation 17).

$$W^{j+1}_{queen} = \begin{cases} W^{j}_{mut,Best} & \text{if } e(W^{l}_{new,Best}) \le e(W^{l}_{queen}) \\ W^{j}_{queen} & \text{Otherwise} \end{cases} \quad (17)$$

Where W^{j+1}_{queen} is the best scheduling/resource allocation policy for the next iteration, ensuring the algorithm moves toward optimal solutions. The mutant vectors are first evaluated using the fitness function, and the best-performing vector (mutant-best) is selected as a candidate policy. This candidate is then compared with the current policy, and if it provides better reward performance, it replaces the existing policy parameters. Finally, the selected mutant vector is used to update the RDQN weights, influencing exploration strategy and improving policy refinement. Table 3 shows the hyper parameter table for DHBM.

Table 3: Hyper parameter table for DHBM

Hyperparameter	**Value**
Learning Rate (α)	0.0005 – 0.005 (DHBM–adaptive)
Discount Factor (γ)	0.90 – 0.99
Exploration Rate (ε)	1.0 → 0.05 (decayed using DHBM)
Batch Size	32 – 128
Replay Buffer Size	10,000 – 50,000
Target Network Update Interval	Every 10–50 episodes
Mutation Coefficients (E1–E3)	0.2 – 0.8
Queen Speed Decay (α_decay)	0.90 – 0.98
Number of Mutant Vectors per Iteration	4

By integrating DHBM with RDQN, the algorithm dynamically tunes the learning process of the reinforcement learning agent, ensuring that scheduling and resource allocation decisions remain effective under real-time variability in patient arrivals, health emergencies, and staff availability. This hybrid mechanism improves both convergence stability and policy adaptability, making it particularly suitable for complex elderly care environments. The pseudocode of the proposed method is given in Algorithm 2.

Algorithm 2: DHBM-RDQN

Input: Preprocessed elderly care dataset, replay buffer
Output: Optimized scheduling policy and reduced patient waiting time
Initialize RDQN parameters θ, target network θ', and replay buffer
Initialize DHBM population (queen + drones) with random policy weights
Evaluate the fitness of each policy → select best as initial queen
While training not converged do
Observe current state s_t
Select action b_t using DHBM-guided exploration (adaptive ε)
Execute action, receive reward r_t and next state s_{t+1}

Store experience tuple (s_t, b_t, r_t, s_{t+1}) in replay buffer
RDQN Update:
Sample mini-batch from replay buffer
compute loss using Equation (9) and update network weights θ

$$J_\theta = \frac{1}{2}\left(Q_\theta(s_t, b_t) - \left(r_t + \gamma \max_{b_j} Q_\theta(s_{t+1}, b_{max})\right)\right)^2$$

Periodically update target network θ'
DHBM Optimization:
Generate mutant vectors using Equations (12–16)

$$W^j_{mutant1} = W^j_{q1} + E_1 \times \left(W^j_{queen} - W^j_{r3}\right) + E_1 \times (W^j_{q3} - W^j_{q4})$$

$$W^j_{mutant2} = W^j_{queen} + E_2 \times \left(W^j_{q1} - W^j_{q2}\right)$$

$$W^j_{mutant3} = W^j_{q1} + E_3 \times \left(W^j_{q2} - W^j_{q3}\right) + E_3 \times (W^j_{q1} - W^j_{q4})$$

$$W^j_{mutant4} = (W^j_{q1} + W^j_{q2} + W^j_{q3})/3 + (\beta_2 - \beta_1)\left(W^j_{q1} - W^j_{q2}\right) + (\beta_3 - \beta_2)(W^j_{q2} - W^j_{q3}) + (\beta_1 - \beta_3)(W^j_{q3} - W^j_{q1})$$

$$\beta_1 = \frac{|e(W^j_{q1})|}{\beta^*}, \beta_2 = \frac{|e(W^j_{q2})|}{\beta^*}, \beta_3 = \frac{|e(W^j_{q3})|}{\beta^*}$$

Compute fitness of all mutant policies
Select best mutant vector → $W_{mut,best}$
If $W_{mut,best}$ improves reward → update queen and RDQN weights
Update exploration rate ε and queen speed $T(l)$ (Eq. 10–11)

$$Prob(C) = \exp(-\Delta(e)/T(l))$$

$$T(l+1) = \alpha \times T(l)$$

End while
Return final optimized policy and performance metrics

4 Experimental result

The implementation details, system configuration, and hyper parameter settings of the proposed DHBM-RDQN method were presented in this section. It highlights the outcomes of the experiments, demonstrating the framework's ability to efficiently allocate resources, optimize staff scheduling, and adapt dynamically to changing elderly care demands.

4.1 System configuration and hyper parameter tuning

The proposed DHBM-RDQN framework was implemented using Python with TensorFlow, PyTorch, and Keras, supported by standard scientific libraries for preprocessing and feature engineering. Model training and evaluation were executed on a GPU-enabled computing environment to ensure efficient convergence and experimentation. Table 4 summarizes the optimized hyper parameters used in the final configuration.

The hyperparameters for DHBM-RDQN were selected through a combination of empirical testing and bio-inspired optimization principles. Baseline values for the RDQN component, such as the number of hidden layers, neurons, activation function, and optimizer, were chosen based on standard reinforcement learning literature to ensure stable learning. The Honeybee Mating (DHBM) layer parameters—population size, number of generations, mating flight length, mutation probability, and crossover probability—were iteratively tuned to enhance exploration, prevent premature convergence, and optimize policy learning under dynamic elderly care conditions. Experimental validation confirmed that these settings provided the best balance between learning stability, convergence speed, and adaptability

Table 4: Hyper parameter setting for DHBM-RDQN

Category	Hyper parameter	Value
RDQN	Hidden Layers	2 fully connected layers
	Neurons per Layer	256, 128
	Activation Function	ReLU
	Optimizer	Adam
	Learning Rate (α)	0.0001 – 0.001
	Discount Factor (γ)	0.95
	Exploration Strategy	ε-greedy
	Initial Exploration Rate (ε₀)	1.0
	Minimum Exploration Rate (ε min)	0.05
DHBM	Learning Rate (α)	0.0005 – 0.005 (DHBM–adaptive)

	Discount Factor (γ)	0.90 – 0.99
	Exploration Rate (ε)	1.0 → 0.05 (decayed using DHBM)
	Batch Size	32 – 128
	Replay Buffer Size	10,000 – 50,000
	Target Network Update Interval	Every 10–50 episodes
	Mutation Coefficients (E1–E3)	0.2 – 0.8
	Queen Speed Decay (α decay)	0.90 – 0.98
	Number of Mutant Vectors per Iteration	4

4.2 Performance evaluation of the proposed framework

Figure 2 shows the variation in patient waiting time across 100 training episodes. The waiting time fluctuates between 15 and 20 minutes, demonstrating dynamic policy adaptation. Spikes and drops indicate learning adjustments based on changing care demands. Overall, the plot shows the model stabilizing around 17–18 minutes, reflecting improved scheduling efficiency.

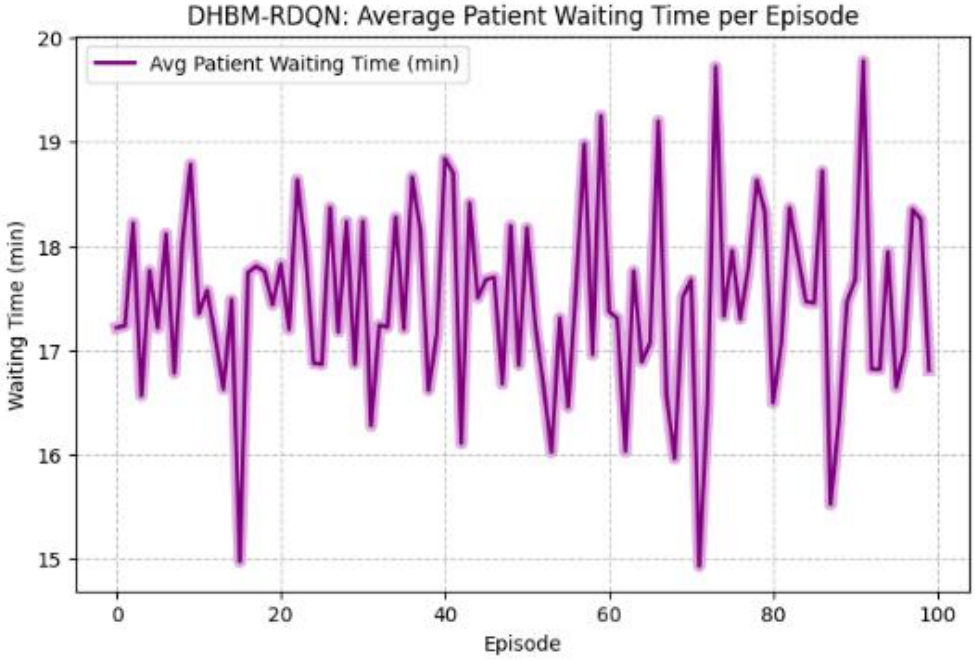


Figure 2: Average patient waiting time for using DHBM-RDQN

Figure 3 illustrates staff overtime fluctuations across the same 100 episodes. Overtime values range between 1.0 and 2.8 hours, showing adaptive workforce utilization. Initial fluctuations gradually reduce, indicating learning stabilization over time. The overall trend settles near 1.8–2.2 hours, reflecting improved workforce balance.

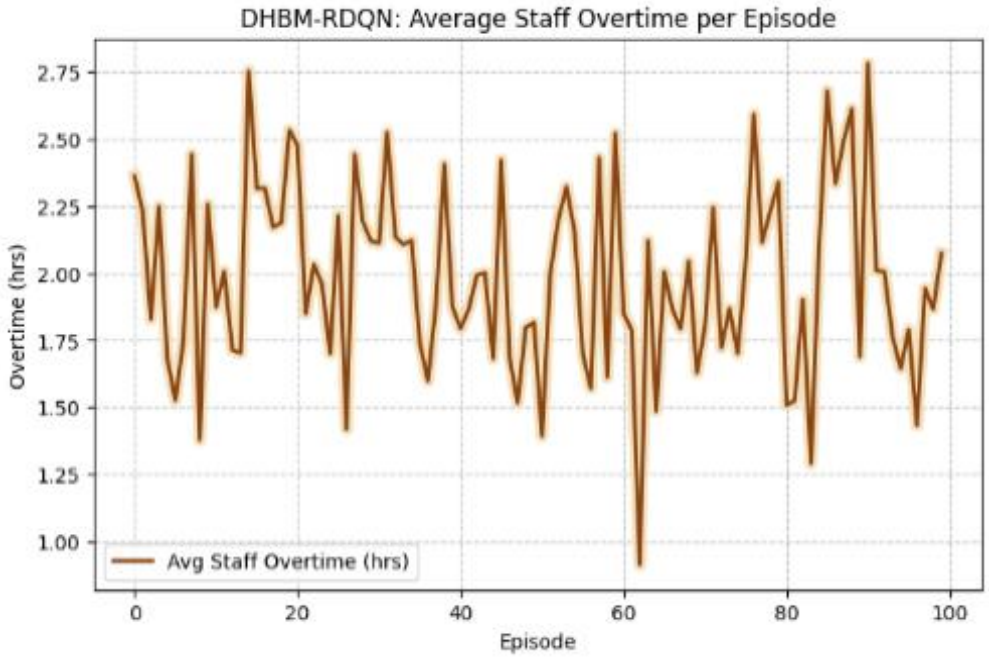


Figure 3: Staff overtime fluctuations for using DHBM-RDQN

Figure 4 displays the reward distribution across 100 training episodes. Reward values increase from approximately 70 to 135, indicating learning progression. The visible upward trend reflects improved decision-making in dynamic resource allocation. Higher reward consistency after episode ~60 indicates stabilization of the trained policy.

This cumulative reward chart highlights learning improvement over time. Values steadily increase from 0 to over 10,000, demonstrating continuous gain in performance. The smooth upward slope indicates consistent learning without policy collapse. By episode 100, the cumulative reward reflects a successfully converged and optimized RL model.

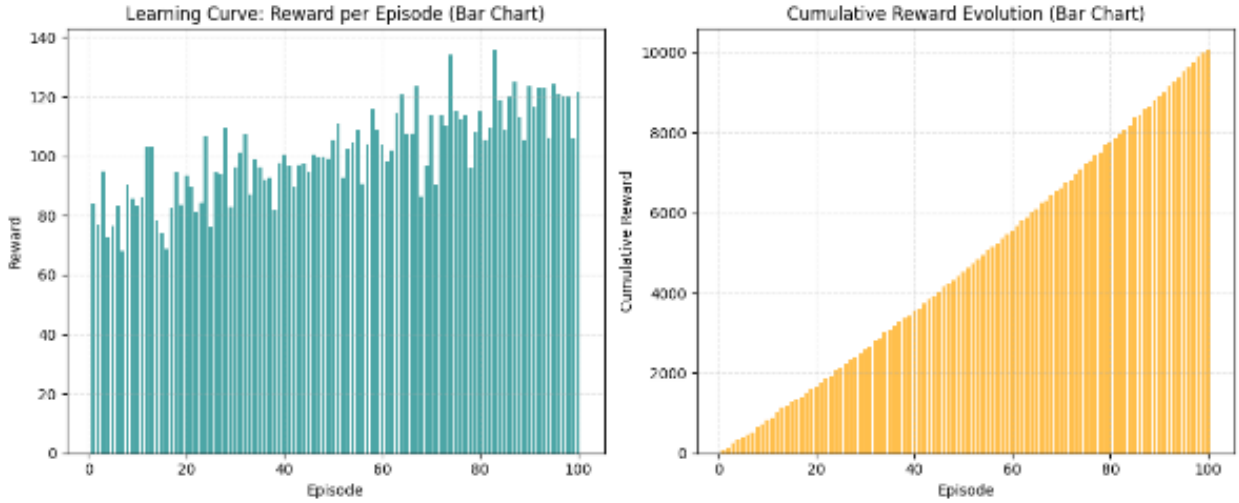


Figure 4: Reward and cumulative reward fluctuations

Figure 5 shows how the proposed DHBM-RDQN responds in a high-demand scenario. The model assigns 80% priority for high patient acuity and 70% allocation when staff availability demonstrating strong adaptability. It also allocates 75% for nurse assignment, 65% for care assistants, and gives 85% priority to urgent task completion, showing improved responsiveness and smarter resource distribution. Overall, DHBM-RDQN demonstrates stronger adaptability and more efficient resource allocation.

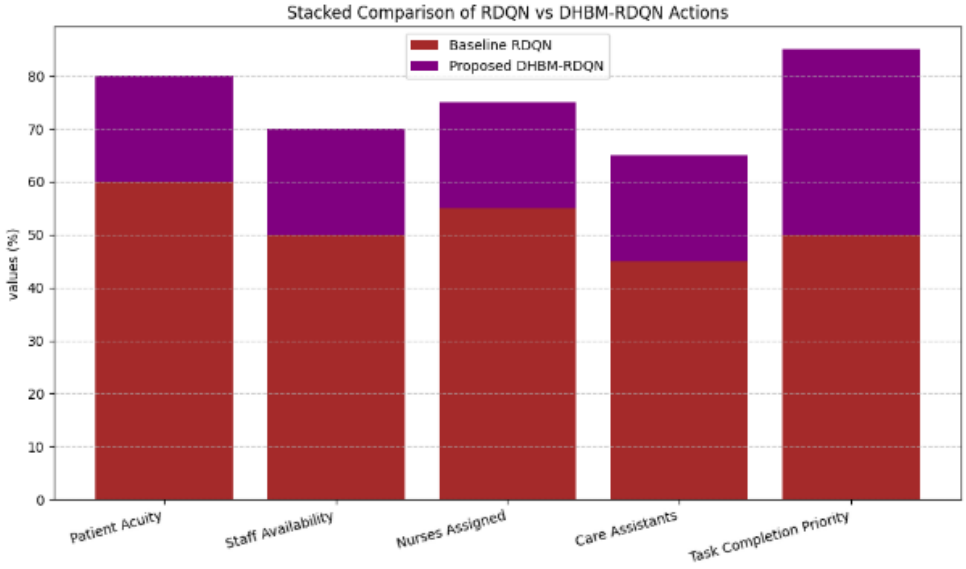


Figure 5: The proposed DHBM-RDQN and Baseline RDQN performance evaluation

The SHAP summary Figure 6 shows the impact of different features on the DHBM-RDQN model's predictions for dynamic elderly care resource allocation. Features like complaint_count, satisfaction_score, and high_acuity_pct have the strongest influence, highlighting the importance of patient feedback and acuity in staffing decisions. Color coding indicates feature values, with red representing high values and blue low values, showing how feature magnitude affects the model output. Operational variables such as avg_hourly_wage_lpn, lpn_headcount, and budget_cap also contribute to resource allocation decisions. Overall, the plot identifies key factors driving the model's dynamic allocation strategy to optimize elderly care quality and efficiency.

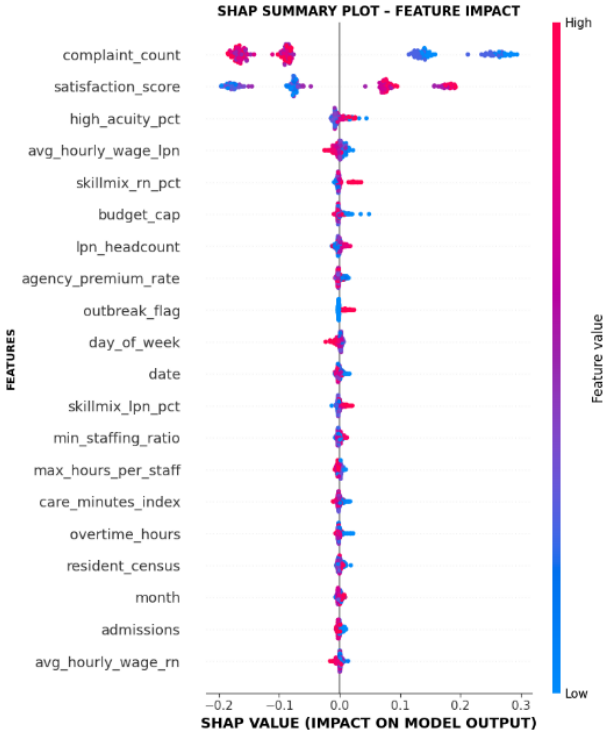


Figure 6: SHAP summary plot for the impact of different features on the DHBM-RDQN model's

The experimental evaluation using five-fold cross-validation demonstrates clear performance differences among DHBM, RDQN, and the hybrid DHBM-RDQN model. The hybrid model consistently achieved the highest accuracy, precision, recall, and F1-score with minimal performance variance, indicating improved generalization and stability. RDQN showed moderate improvement over DHBM, while DHBM exhibited the lowest performance but remained stable. Statistical significance analysis confirmed that the performance gains of DHBM-RDQN were not random and were highly significant ($p < 0.01$).

Table 5: 5-fold cross validation

Model	Accuracy (Mean ± SD)	95% CI	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)	Significance
DHBM	0.87 ± 0.012	[0.85–0.89]	0.85 ± 0.015	0.84 ± 0.017	0.85 ± 0.013	Baseline
RDQN	0.90 ± 0.010	[0.89–0.92]	0.89 ± 0.012	0.88 ± 0.014	0.88 ± 0.011	↑ Significant vs DHBM ($p < 0.05$)
DHBM-RDQN (Proposed)	0.95 ± 0.008	[0.94–0.97]	0.94 ± 0.010	0.95 ± 0.009	0.95 ± 0.008	↑↑ Highly Significant vs DHBM & RDQN ($p < 0.01$)

The Table 5 shows that the proposed DHBM-RDQN model achieved the highest overall performance, reaching an accuracy of 0.95 ± 0.008, precision of 0.94 ± 0.010, recall of 0.95 ± 0.009, and an F1-score of 0.95 ± 0.008, outperforming both DHBM and RDQN. The confidence interval (0.94–0.97) and low standard deviation highlight strong model reliability. In comparison, DHBM achieved the lowest scores, while RDQN demonstrated moderate improvements with statistically significant gains relative to DHBM.

4.3 Performance comparison with existing method

To validate the effectiveness of the proposed DHBM-RDQN method, its performance was compared against three benchmark approaches: a Time Series Analysis framework [23] (which integrates hybrid time series forecasting with mathematical programming using optimization-based hyperparameter tuning), a traditional RNN model [23] and an LSTM-based architecture [23].. The evaluation was carried out across four key metrics, such as accuracy, resource efficiency, response time, and adaptability score, to comprehensively assess predictive reliability, optimal resource allocation, real-time responsiveness, and adaptability to dynamic elderly care environments.

Adaptability score: Adaptability score measures how well the system adjusts resource allocation when patient conditions or staffing levels change. It reflects the ratio of successful adaptive decisions to total decision events. Equation (18) shows the adaptability score mathematical representation:

$$\text{Adaptability score} = \frac{\text{Adaptive decision corretly executed}}{\text{Total Adaptive Descision Events}} \quad (18)$$

Response time: Response time indicates how quickly the model reacts to changes in care demand and reallocates resources. It is measured as the average time taken by the system to output a decision after receiving new state input. The formula of response time is shown in Equation (19):

$$\text{Response time} = \sum_{i=1}^{n} \text{Decision Output time} - \text{Input timestamp}_n \quad (19)$$

Resource efficiency: Resource efficiency measures how effectively available staff and services are allocated without overuse or underutilization. It compares the optimal resource usage to the actual usage determined by the model. The mathematical representation of resource efficiency is shown in Equation (20):

$$\text{Resource Efficiency} = \frac{\text{Optimal Resource Utilization}}{\text{Actual Resource Utilization}} \times 100\% \quad (20)$$

Accuracy: Accuracy reflects how many allocation decisions made by the system match the expected or expert-approved optimal allocation.

It is computed as the ratio of correct decisions to total decisions. Accuracy formula is shown in Equation (21):

$$\text{Accuracy} = \frac{\text{Correct Allocation Decisions}}{\text{Total Allocation Decisions}} \times 100\% \quad (21)$$

Table 6 gives the comparative evaluation of the proposed method with baseline approaches. Figure 7 illustrates the accuracy comparison among the proposed DHBM-RDQN model and existing approaches, including LSTM-based, Traditional RNN, and the Time Series Analysis framework. As seen, the DHBM-RDQN achieves the highest accuracy (96.5%), outperforming all baselines, which highlights its effectiveness in capturing complex temporal dependencies and decision-making dynamics. Figure 7 also shows the comparison of resource efficiency across the models. The DHBM-RDQN demonstrates superior efficiency (90.5%) compared to LSTM and Traditional RNN, while also superior to the Time Series Analysis framework. This indicates that the proposed method utilizes computational resources more effectively, making it highly suitable for real-time and large-scale applications.

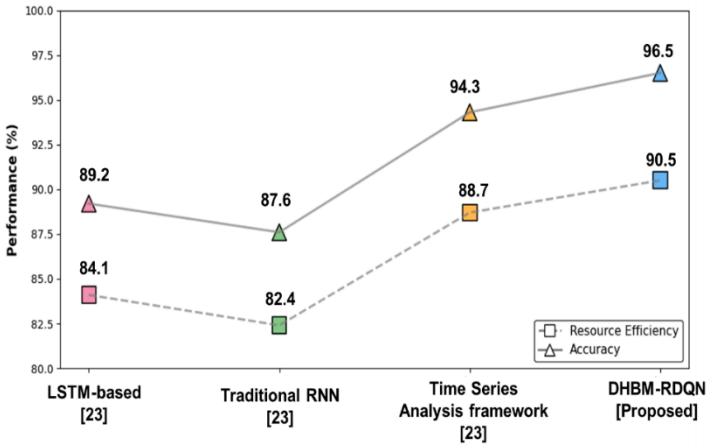


Figure 7: Accuracy and Resource efficiency performance of proposed vs. existing models

Figure 8 presents a joint comparison of response time and adaptability score. The DHBM-RDQN achieves the lowest response time (1.0s) and the highest adaptability score, proving its robustness and faster decision-making capabilities under dynamic conditions. Compared to traditional baselines, LSTM and Traditional RNN, and also the Time Series Analysis framework, this improvement highlights the adaptability and efficiency of the proposed approach in rapidly changing environments.

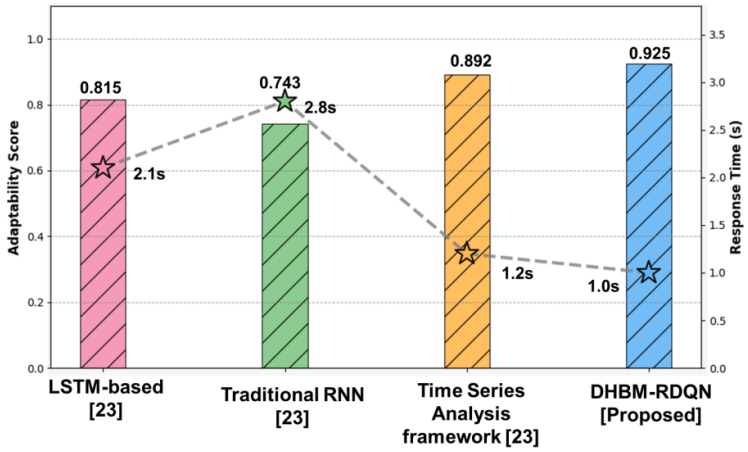


Figure 8: Score comparison of response time and adaptability score

Table 6: Performance comparison of DHBM-RDQN with existing methods

Method	Adaptability Score	Response Time	Resource Efficiency	Accuracy
LSTM-based [23]	0.815	2.1s	84.1 %	89.2 %
Traditional RNN [23]	0.743	2.8s	82.4 %	87.6 %
Time Series Analysis framework [23]	0.892	1.2s	88.7 %	94.3 %
DHBM-RDQN[Proposed]	0.925	1.0s	90.5 %	96.5 %

Existing methods of managing elderly care resources, including Traditional RNN, LSTM-based models, and Time Series Analysis frameworks [23], have multiple limitations. Traditional RNNs fail to learn long-term dependencies effectively due to vanishing gradients, LSTM models learn temporal patterns more effectively but require a lot of computational power to train them for a long time, and Time Series Analysis frameworks rely on a fixed mathematical assumption(s) and optimization heuristics that can't adapt to fluctuations in resource availability or patient demand as they happen. The proposed DHBM-RDQN method enables Deep RL with a high-dimensional state-action mapping with a DHBM method and simultaneously resolves all of the limitations described above. The RDQN component learns a high-dimensional state-action mapping without manual

engineering of the features, whereas DHBM guarantees exploration without premature convergence or exploitation to provide optimal allocations when uncertainty exists. As a result, the DHBM-RDQN method improved accuracy and resource efficiency, allowed for faster response, and accommodated greater fluctuations in resource availability. Taking all the improvements into consideration, DHBM-RDQN is significantly better than traditional methods of resource allocation in highly complex situations in the elderly care sector.

4.4 Discussion

The LSTM [23] models are suitable in time-sequence prediction but fail at real-time decision-making in dynamic settings such as the case of providing elderly care. They need regular patterns to be able to generalize and fail when staff or patient demand changes in a random fashion. LSTM does not have the ability to automatically modify resource allocation policies and relies on pre trained hard coded rules. Consequently, flexibility and real-time optimization are minimal.

Older RNNs [23] struggle to deal with long-term dependency dynamics, and thus it deteriorates their performance when dealing with continuous and multi-shift operational data. They are also likely to lose their training when faced with complex decision environments and a long planning horizon. RNNs do not require interaction with the environment and can therefore not learn policies optimally or on the feedback of the environment, as they are not reinforcement-based. This restricts their applicability in dynamic staffing and allocation of resources situations.

The time series forecasting [23] methods can be used to predict the trends in demand and cannot make optimal decisions on their own based on the projected states. They are run on fixed assumptions and they do not have systems to accommodate unforeseen occurrences like sudden shortages of staff or emergencies in patients. Such practices lack a reward-based learning mechanism to review and filter down decisions. They, therefore, offer resources but not automated and adaptive resources allocation.

To address these limitations, the proposed Dynamic Honeybee Mating–tuned Resource-based Deep Q-Network (DHBM-RDQN) introduces a reinforcement learning-driven framework capable of making real-time adaptive decisions. The model continuously learns from environment feedback and dynamically allocates care resources based on changing patient needs and staffing conditions. The honeybee mating optimization mechanism further enhances learning stability by tuning hyper parameters automatically, improving exploration and preventing premature convergence. As a result, the proposed method achieves higher accuracy, better adaptability, and improved resource efficiency compared to existing predictive and non-learning-based approaches.

The proposed DHBM-RDQN framework is promising, The DHBM-RDQN framework dynamically adapts to unexpected staff shortages by reallocating available caregivers based on real-time patient demand and acuity levels. Its mutation-driven exploration generates alternative scheduling policies, ensuring service continuity and minimizing patient waiting times. This adaptive policy learning allows the system to maintain efficiency even under sudden workforce fluctuations.

5 Conclusion

A novel approach for dynamic allocation of elderly care service resources using RL, specifically the DHBM-RDQN method, was presented in this research. The proposed dynamic allocation framework integrated 24 months of operational data from 3 large-scale elderly care facilities, representing over 30,000 service records, including staff schedules, patient health characteristics, emergency event documentation, and care quality. It did so in a theoretically guided way to model the complexities of resource management in the real-world. The operational data was pre-processed so it was standardized (Z-score normalization) and could include missing data (imputation). Feature extraction methods were then applied to the data to derive latent patterns and temporal dynamics as part of the state-space representation of an MDP. Experimental results showed that the RL-based framework achieved 96.5% accuracy, 90.5% resource efficiency, a 1.0 s average response time, and an adaptability score of 0.925, significantly outperforming traditional methods. In conclusion, this research provides a valuable and generalizable approach to dynamically allocating care service resources through a RL framework to enable efficient, data-driven decision making for elderly care management. It also incorporated an MDP model, pre-processing, and a feature extraction method that help establish a solid empirical basis for the intelligent, scalable and practical optimization of resources in long-term care settings.

Limitations and Future Scope: The proposed DHBM-RDQN framework, while effective, has certain limitations. Its generalizability to other elderly care datasets or facilities may be constrained due to differences in patient demographics, staffing patterns, or operational practices. The RL model assumes relative stationarity in state transitions and may not fully capture delayed effects of actions on care outcomes. Additionally, potential biases in the dataset, such as variations in socioeconomic status or facility resources, could influence the learned policies and affect fairness in resource allocation. Future research can integrate real-time IoT-based monitoring, explore multi-agent reinforcement learning for coordinated care, and scale the approach to broader regional or national elderly care networks.

Author Contributions

Jiang. Projectadministration& Conceptualization& Visualization&Experiment operation and execution& Review and revision

References

[1] Tsang, Y.P., Wu, C.H., Leung, P.P., Ip, W.H. and Ching, W.K., 2021. Blockchain-IoT-driven nursing workforce planning for effective long-term care management in nursing homes. Journal of healthcare engineering, 2021(1), p.9974059. https://doi.org/10.1155/2021/9974059

[2] Saaiman, T., Filmalter, C.J. and Heyns, T., 2021. Important factors for planning nurse staffing in the emergency department: a consensus study. International Emergency Nursing, 56, p.100979. http://dx.doi.org/10.1016/j.ienj.2021.100979

[3] Ou, X., 2025. Dynamic Constraint-Aware Particle Swarm Optimization for Resource Allocation in Logistics and Transportation. Informatica, 49(29). https://doi.org/10.31449/inf.v49i29.9209

[4] Secundo, G., Shams, S.R. and Nucci, F., 2021. Digital technologies and collective intelligence for healthcare ecosystem: Optimizing Internet of Things adoption for pandemic management. Journal of Business Research, 131, pp.563-572. https://doi.org/10.1016/j.jbusres.2021.01.034

[5] Lee, D. and Yoon, S.N., 2021. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International journal of environmental research and public health, 18(1), p.271. https://doi.org/10.3390/ijerph18010271

[6] Yang, J., Luo, B., Zhao, C. and Zhang, H., 2022. Artificial intelligence healthcare service resources adoption by medical institutions based on TOE framework. Digital Health, 8, p.20552076221126034. DOI: 10.1177/20552076221126034

[7] Ali, H., 2022. Reinforcement learning in healthcare: optimizing treatment strategies, dynamic resource allocation, and adaptive clinical decision-making. Int J Comput Appl Technol Res, 11(3), pp.88-104. DOI: 10.7753/IJCATR1103.1007

[8] Wu, Q., Han, J., Yan, Y., Kuo, Y.H. and Shen, Z.J.M., 2025. Reinforcement learning for healthcare operations management: methodological framework, recent developments, and future research directions. Health Care Management Science, pp.1-36. https://doi.org/10.1007/s10729-025-09699-6

[9] Yazdani, M., Shahriari, S. and Haghani, M., 2025. Real-time decision support model for logistics of emergency patient transfers from hospitals via an integrated optimisation and machine learning approach. Progress in Disaster Science, 25, p.100397. https://doi.org/10.1016/j.pdisas.2024.100397

[10] Thomas, J., 2024. Optimizing nurse scheduling: a supply chain approach for healthcare institutions. arXiv preprint arXiv:2407.11195. https://doi.org/10.48550/arXiv.2407.11195

[11] Talaat, F.M., 2022. Effective deep Q-networks (EDQN) strategy for resource allocation based on optimized reinforcement learning algorithm. Multimedia Tools and Applications, 81(28), pp.39945-39961.https://doi.org/10.1007/s11042-022-13000-0

[12] Fischer, G.S., da Rosa Righi, R., de Oliveira Ramos, G., da Costa, C.A. and Rodrigues, J.J., 2020. ElHealth: Using Internet of Things and data prediction for elastic management of human resources in smart hospitals. Engineering Applications of Artificial Intelligence, 87, p.103285. https://doi.org/10.1016/j.engappai.2019.103285

[13] Song, Y. and Wu, R., 2021. Analysing human-computer interaction behaviour in human resource management system based on artificial intelligence technology. Knowledge Management Research & Practice, pp.1-10. https://doi.org/10.1080/14778238.2021.1955630

[14] Yinusa, A. and Faezipour, M., 2023. Optimizing healthcare delivery: A model for staffing, patient assignment, and resource allocation. Applied System Innovation, 6(5), p.78. https://doi.org/10.3390/asi6050078

[15] Ghayoora, F. and Rahmania, D., Improving Healthcare Service Quality: A System Dynamics Approach to Managing Visit Times and Reducing Error Rates. https://jstinp.um.ac.ir/article_46620.html

[16] Olya, M.H., Badri, H., Teimoori, S. and Yang, K., 2022. An integrated deep learning and stochastic optimization approach for resource management in team-based healthcare systems. Expert Systems with Applications, 187, p.115924. http://dx.doi.org/10.1016/j.eswa.2021.115924

[17] Aslan, M. and Toros, E., 2025. Machine Learning in Optimising Nursing Care Delivery Models: An Empirical Analysis of Hospital Wards. Journal of Evaluation in Clinical Practice, 31(1), p.e70001. https://doi.org/10.1111/jep.70001

[18] Leung, F., Lau, Y.C., Law, M. and Djeng, S.K., 2022. Artificial intelligence and end user tools to develop a nurse duty roster scheduling system. International Journal of Nursing Sciences, 9(3), pp.373-377. https://doi.org/10.1016/j.ijnss.2022.06.013

[19] Pranata, A. and Yudhantara, R., 2023. Strategic Human Resource Allocation in Healthcare Institutions Using AI-Enabled Workforce Analytics

and Predictive Modeling. International Journal of Theoretical, Computational, and Applied Multidisciplinary Sciences, 7(12), pp.1-24.

[20] Huang, X., 2025. Approximate SARSA Algorithm for Dimensionality-Challenged Resource Allocation Optimization in MIMO Communication Systems. Informatica, 49(8). https://doi.org/10.31449/inf.v49i8.7251

[21] Safdar, K.A., Emrouznejad, A. and Dey, P.K., 2020. An optimized queue management system to improve patient flow in the absence of appointment system. International Journal of Health Care Quality Assurance, 33(7/8), pp.477-494. https://doi.org/10.1108/IJHCQA-03-2020-0052

[22] Lee, S. and Lee, Y.H., 2020, March. Improving emergency department efficiency by patient scheduling using deep reinforcement learning. In Healthcare (Vol. 8, No. 2, p. 77). MDPI. http://dx.doi.org/10.3390/healthcare8020077

[23] Ma, D. and Ling, Z., 2024. Optimization of Nursing Staff Allocation in Elderly Care Institutions: A Time Series Data Analysis Approach. Annals of Applied Sciences, 5(1).

Appendix

Abbreviation	Full form
ML	machine learning
NLP	natural language processing
QOC	quality of care
AI	artificial intelligence
DL	deep learning
RL	Reinforcement learning
MDP	Markov Decision Process
ERAS	effective resource allocation strategy
ARU	average resource utilization
PSO	particle swarm optimization
ElHealth	Elastic allocation of human resources in healthcare environments
HRM	human resource management
ED	Emergency Department
IoT	Internet of Things
BPNN	Back propagation neural network
LOO	Leave-One-Out
MILP	mixed-integer linear programming
RF	Random Forest
SERVQUAL	Service Quality
DEA	Data Envelopment Analysis
VBA	Visual Basic for Applications
ED	Emergency departments
DQN	Deep Q-networks
BP	blood pressure
MVI	Missing Value Imputation
HR	heart rate
ICA	Independent Component Analysis
DWT	Discrete Wavelet Transform
WT	Wavelet Transform
HBMO	Honey bee mating optimization
traditional RNN	traditional Recurrent neural network
LSTM	Long short-term memory
NCD	nursing care delivery
DNN	Deep Neural Network

Adaptive Multidimensional Fusion Network with Dynamic Decision Trees for Financial Market Trend Forecasting

Sheng Wang[1], Jing Wu[2*]
[1]College of Economics and Management, Jiaozuo University, Jiaozuo 454000, China
[2]College of Continuing Education, Jiaozuo University, Jiaozuo 454000, China
E-mail: WangSheng_ws@outlook.com, wujing126@hotmail.com
[*]Corresponding author

Keywords: deep neural network, financial market trend prediction, adaptive multi-dimensional fusion network, data fusion, dynamic adaptive mechanism

Received: August 11, 2025

This study proposes an Adaptive Multidimensional Fusion Network (AMFN) for financial market trend forecasting. The model integrates heterogeneous data sources through a multidimensional data fusion module, combining historical price and volume data with external information such as macroeconomic indicators and sentiment indices. An adaptive temporal processing module is employed to model time-varying dependencies and regime shifts in market behavior, while a dynamic decision-tree prediction module captures nonlinear patterns in the fused representations. Experiments are conducted on multiple financial datasets, including the S&P 500 Index, China A-share market, and Gold Futures, using a rolling time-window evaluation to avoid information leakage. The AMFN model achieves lower MSE and MAE and higher R^2 than traditional SVM and LSTM baselines, with up to 24.4% relative improvement in forecasting accuracy. These results demonstrate that AMFN provides interpretable, stable, and robust trend predictions across diverse market environments.

Povzetek: Študija predlaga adaptivno večdimenzionalno fuzijsko nevronsko mrežo (AMFN), ki z združevanjem tržnih zgodovinskih podatkov in zunanjih kazalnikov izboljša natančnost ter robustnost napovedovanja finančnih trendov v različnih trgih.

1 Introduction

Financial markets are highly volatile and influenced by sudden external shocks, making trend forecasting difficult despite abundant historical data [1]. Events such as the 2008 global financial crisis highlighted how rapidly market conditions can shift and revealed the limitations of traditional predictive models [2][3]. As market dynamics increasingly integrate economic indicators, global events, and social sentiment, the complexity of price formation continues to grow [4].

Deep neural networks (DNNs), including RNNs and LSTMs, improve pattern recognition and temporal modeling compared with linear and classical time-series methods [5][6]. However, challenges remain, including overfitting, limited long-term prediction robustness, and insufficient use of external information [7][8]. Moreover, market movements are often driven by non-price factors such as geopolitical events and sentiment.

This study addresses these issues by integrating heterogeneous data sources to enhance predictive accuracy, contributing to more informed investment decisions and advancing deep learning–based financial forecasting.

Based on these gaps, we formulate the following research questions:

(1) Can integrating external sentiment and macroeconomic indicators with historical time-series data through adaptive fusion improve predictive accuracy?

(2) How does AMFN perform relative to deep and hybrid models under high-volatility market conditions, where regime shifts and nonlinear behaviors intensify?

These questions guide the model design, experimental setup, and comparative evaluation presented in this study.

2 Lliterature review

2.1 Application of deep neural networks in financial market prediction

Financial market forecasting remains challenging due to the nonlinear, volatile, and dynamic nature of market data. Traditional statistical models such as linear regression and time series analysis offer limited capability in capturing these complex patterns. With advances in machine learning, deep neural networks (DNNs) have demonstrated strong potential in financial trend prediction [9]. Architectures such as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) effectively model sequential dependencies and improve prediction accuracy for stock prices and market indices [10]. LSTMs further address gradient vanishing and enhance long-term pattern learning [11], outperforming traditional models in trend recognition [12]. However,

deep learning models still suffer from overfitting, limited interpretability, and sensitivity to parameter tuning [13], which restricts their deployment in high-stakes financial decision-making. To mitigate these challenges, recent studies incorporate sentiment signals, economic indicators, and attention mechanisms to enhance feature representation and improve model transparency and robustness in complex financial environments [14–18].

2.2 The role of external data sources in financial forecasting

Recent studies increasingly emphasize that historical price data alone cannot fully characterize market dynamics. External data sources such as news, social media sentiment, and macroeconomic indicators have become essential for forecasting [19]. Sentiment extracted from platforms like financial news feeds and Twitter can act as a leading indicator of market direction, particularly during periods of uncertainty [20–22]. In parallel, macroeconomic variables such as GDP growth and inflation remain key drivers of structural market behavior and can help anticipate significant trend shifts or crisis conditions . Hybrid models that fuse price data with sentiment and economic data show improved ability to capture both short-term fluctuations and long-term market cycles. However, integrating heterogeneous data introduces challenges in representation alignment, noise filtering, and adaptive weighting. To address this, attention-based fusion mechanisms are being explored to dynamically adjust input importance based on market states, improving prediction reliability under varying volatility conditions.

2.3 Challenges and future development directions of deep neural networks in financial forecasting

Despite strong predictive ability, deep neural networks face several limitations in financial forecasting. Markets are influenced by unpredictable events such as geopolitical shocks and pandemics, which historical patterns cannot fully capture. This reduces model generalization and performance stability. Additionally, the "black-box" nature of deep neural networks limits interpretability, which is critical for financial decision-making . Explainable AI techniques such as LIME and SHAP are being explored to clarify feature contributions, though widespread application remains difficult. Another emerging direction is deep reinforcement learning, enabling adaptive strategy optimization through interaction with market environments; however, balancing exploration and exploitation remains challenging. Computational demands also increase with dataset scale, motivating research into pruning, quantization, and transfer learning for efficiency. While prior studies such as Zhang et al. [2] and Sawhney et al. [3] integrate external information, the AMFN model advances the field by incorporating automated data-source weighting and dynamic decision-tree–based nonlinear prediction. Future work should compare AMFN more directly with Transformer-based and graph-based models to further clarify its advantages.

To provide a structured comparison, we summarize representative forecasting models in Table 1. The comparison covers model type, primary data inputs, evaluation metrics, and reported performance. Classical SVM and LSTM approaches generally rely on historical prices only, while GCN- and Transformer-based models incorporate relational or long-range dependency structures. However, few models explicitly combine external sentiment and macroeconomic variables with adaptive weighting. This gap motivates the AMFN framework, which unifies heterogeneous data, dynamic temporal adaptation, and nonlinear decision-based prediction.

Table 1: Comparative summary of related methods

Model	Data Sources	Key Technique	Evaluation Metrics	Reported Performance
SVM	Price series	Static kernel mapping	MSE, MAE	Moderate, low adaptability
LSTM	Price + volume	Sequential memory	MSE, Directional accuracy	Strong short-term tracking
GCN	Price + correlation graph	Graph structural learning	R^2, Accuracy	Sensitive to graph quality
Transformer	Price + temporal embeddings	Long-range attention	MSE, Accuracy	High capacity, risk of overfitting
AMFN (proposed)	Price + macro + sentiment	Fusion + adaptive temporal + dynamic tree	MSE, MAE, R^2, Accuracy	Highest robustness across markets

3 Methodology

3.1 Innovation model design and theoretical basis

Market features x_t and external features z_t are fused by a gated layer to form h_t . An adaptive recurrent block produces s_t with regime-aware gating. A differentiable dynamic-tree head maps s_t to the target $\hat{y}_t$. We train with Adam, lr 1e-3, batch 32, dropout 0.5, early stopping on validation MSE with patience 10. All preprocessing is fit on train folds only. This study proposes a new deep learning architecture - Adaptive Multidimensional Fusion

Network (AMFN) for Financial Market Trend Forecasting. The innovation of this model is reflected in its unique data fusion method and dynamic adaptive mechanism, which can effectively process the time series data and external influencing factors of the financial market and improve the accuracy of trend forecasting. Compared with traditional neural network methods, AMFN does not rely on a single network architecture, but optimizes the forecast results through the interaction and collaboration of multiple modules.

When building the model, we first made a clear assumption: the behavior of the financial market is not only affected by historical data (such as stock prices, trading volumes, etc.), but also by external information (such as news sentiment, macroeconomic indicators, etc.). Therefore, the model needs to be able to process historical market data and external data at the same time and optimize the relationship between them through adaptive mechanisms. The AMFN model consists of three main modules:

1. Multidimensional data fusion module: responsible for fusing historical market data with external data to form a unified feature representation.

2. Adaptive time series processing module: It is specially designed to process the time series characteristics of financial market data and adapt to market changes through adaptive adjustments.

3. Non-linear trend prediction module: The fused features are non-linearly modeled through a dynamic decision tree algorithm to generate market trend predictions.

In the data input processing, we define the input as the historical market data $X = \{X_1, X_2, \ldots, X_T\}$ where XT represents the data at time ttt. Additionally, external data is represented as $Z = \{Z_1, Z_2, \ldots, Z_T\}$, where ZT includes market-related macroeconomic indicators and sentiment analysis results derived from news sources.

This data is processed and fed into the model to generate a unified feature vector, which combines both historical market data and external factors to create a comprehensive representation of the market at each time step.

$$\tilde{x}_t = f_1(x_t, z_t) \tag{1}$$

In Equation (1), f_1 is a fusion function, which converts data from different sources (market data and external data) into feature vectors of the same dimension. Then, we perform time series processing on the features through the adaptive time series module in Equation (2).

$$\hat{x}_t = f_2(\tilde{x}_t) \tag{2}$$

In Equation (2), f_2 is an adaptive time series processing function, and its design takes into account the dynamics of time series data. Finally, the fused features are processed by the nonlinear trend prediction module to obtain the predicted value of Equation (3).

$$\hat{y}_t = f_3(\hat{x}_t) \tag{3}$$

In Equation (3), f_3 it is a nonlinear trend prediction function, which is implemented through a dynamic decision tree algorithm.

The prediction of market trends, denoted as y_t, refers to the predicted directional movement of the market, typically representing whether the market will experience an upward or downward trend in the upcoming period. The trend is defined based on the relative change in the market value, calculated as the percentage change between consecutive time points.

The functions f_1, f_2, and f_3 are the core components of the AMFN model. These functions are defined as follows:

f_1: Multidimensional data fusion function, which combines historical market data (prices, volumes) with external data (sentiment indices, economic indicators) into a unified feature representation.

f_2: Adaptive time series processing function, which adjusts the model's parameters dynamically based on the temporal dependencies of market data. This function ensures that the model adapts to changing market conditions.

f_3 : Nonlinear trend prediction function, implemented through a dynamic decision tree algorithm, which makes predictions based on the processed data features.

Additionally, the loss function is defined as the weighted sum of the prediction error for both market data and external variables. The predicted external data, z^t, represents the forecasted value of external factors (such as sentiment indices) at time ttt. This value is obtained through a separate predictive model that estimates external factors based on historical data and trends.

The dynamic decision tree algorithm (used in f_3) is based on the recursive partitioning method, where the decision tree's splitting criteria are dynamically adjusted according to the market's volatility and the external data inputs. A more detailed explanation of this algorithm and its integration into the model is provided in.

3.2 Component collaboration and model training

The core of the AMFN model lies in the collaborative work of its adaptive data flow mechanism and multidimensional data fusion module. In the financial market, different data sources have different importance and timeliness. How to adjust the flow and processing of data according to these characteristics is an important challenge in model design. To solve this problem, this model introduces a dynamic weight allocation mechanism, so that during the training process, the model can automatically adjust the weight according to the contribution of each data source to optimize the prediction performance.

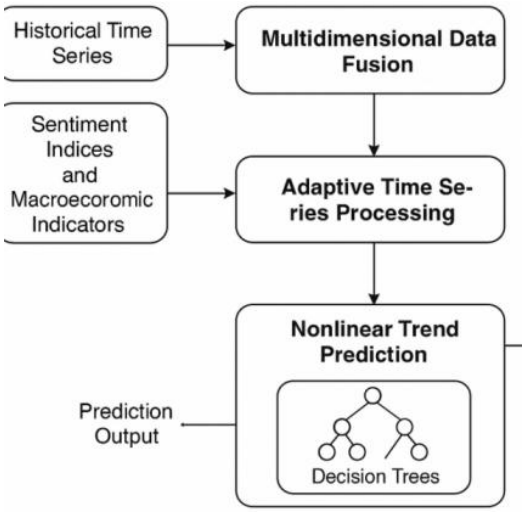


Figure 1: Model structure diagram

Figure 1 now aligns blocks with Equations (1)–(4) and the tree head with Equation (7), clarifying the end-to-end pipeline and implementation. To enhance the reproducibility of our model, we provide an overview schematic of the AMFN architecture in Figure 1. This diagram illustrates the key components of the model, including the multidimensional data fusion module, the adaptive time series processing module, and the nonlinear trend prediction module. Each equation is directly aligned with the corresponding component shown in the diagram. Equation (1) represents the data fusion operation, which combines historical time series data with external variables such as sentiment indices and macroeconomic indicators. Equation (2) corresponds to the time series processing mechanism, allowing the model to adjust dynamically based on recent input trends. Equation (3) captures the nonlinear prediction output generated by the decision tree algorithm. Together, these components define the end-to-end workflow of AMFN and make the model structure transparent and reproducible for future research.

In the data fusion process, the weight of each input feature is set to α_i, which represents i the importance of the data source (such as market historical data or external sentiment data). The input feature fusion at each moment is shown in Equation (4).

$$\tilde{x}_{t'} = \sum_{i=1}^{N} \alpha_i \cdot \tilde{x}_{t,i} \tag{4}$$

In Equation (4), N is the number of data sources, $\tilde{x}_{t,i}$ From the data source i During the training process, the weight coefficients are dynamically adjusted through the back propagation algorithm. α_i, so that the most effective data source has a greater weight in the prediction.

The model is trained by minimizing the prediction error. The objective function is set as the loss function, and Equation (5) includes the weighted sum of the prediction error and the external data error.

$$L = \sum_{t=1}^{T} \left[\lambda_1 \| y_t - \hat{y}_t \|^2 + \lambda_2 \sum_{i=1}^{N} \alpha_i \| z_{t,i} - \hat{z}_{t,i} \|^2 \right] \tag{5}$$

In Equation (5), y_t For the real market trend, $\hat{y}_t$ is the trend predicted by the model, $z_{t,i}$ For external data sources i The true value of $\hat{z}_{t,i}$ is the predicted value, λ_1 and λ_2 is a weighting coefficient used to balance the impact of the two errors. The optimization of the loss function in Equation (6) is achieved by the gradient descent method.

$$\theta_{new} = \theta_{old} - \eta \cdot \frac{\partial L}{\partial \theta} \tag{6}$$

In Equation (6), η is the learning rate, θ are model parameters, $\frac{\partial L}{\partial \theta}$ is the gradient of the loss function with respect to the parameters.

The collaboration of the model's components is fundamental to its performance. The AMFN model consists of three main modules: multidimensional data fusion, adaptive time series processing, and nonlinear trend prediction. The training process is structured as follows:

First, the model is initialized with random weights and trained using the historical data and external data, minimizing the loss function (defined in Equation 5) during each iteration.

The training uses stochastic gradient descent (SGD) with a learning rate of 0.001 and a batch size of 32.

During the training process, we also incorporate dropout with a rate of 0.5 to avoid overfitting.

Each component of the model interacts with the others, allowing the network to learn both temporal dependencies in the market data and the importance of external variables in predicting market trends.

3.3 Nonlinear trend prediction and dynamic decision tree

The dynamic decision tree adapts splitting rules based on real-time market volatility. Let s_t be the temporal state. At each node, features are evaluated using gain score G(f), and the highest-gain feature is selected. Fusion weights α_t are learned through a softmax gating layer $\alpha_t = \text{softmax}(W[h_t; s_t])$. Path weights in the final prediction are optimized end-to-end using backpropagation.

Algorithm:

for each node j:

compute G(f_j) for candidate splits

select f* = argmax G(f_j)

split data using f*update path weights π via gradient descent

In the core module of the model, the nonlinear trend prediction module, we innovatively introduced a dynamic decision tree algorithm. The traditional decision tree algorithm constructs the tree through a fixed splitting rule, while the dynamic decision tree of this model responds to

market changes by adaptively adjusting the splitting rule and the depth of the tree. Each layer of the decision tree automatically selects the most appropriate feature for splitting according to the current market status, making the construction of each tree flexible and adaptable.

Set the output of the decision tree to $\hat{y}_t$, whose decision process is based on the input features $\hat{x}_t$ The decision tree construction process follows the "maximum gain" criterion of Equation (7), splitting at each node based on the gain measure of feature selection.

$$\Delta G = G(\text{parent}) - \sum_{i=1}^{N} G(\text{child}_i) \tag{7}$$

In Equation (7), $G(\text{parent})$ and $G(\text{child}_i)$ Represent the gains of the parent node and child node respectively, N is the number of child nodes of the current node. Each time the split is performed, the feature with the largest gain is selected for further splitting until the stopping condition is met (such as the depth of the tree or the number of nodes reaches the set threshold).

In order to enhance the nonlinear expression ability of the model, we introduced a multi-path fusion mechanism. When building a decision tree, each path can independently generate a prediction value. The final prediction value is obtained by weighted average of the outputs of the paths in Equation (8).

$$\hat{y}_t = \sum_{p=1}^{P} \beta_p \cdot y_{t,p} \tag{8}$$

In Equation (8), P is the number of decision tree paths, β_p For path p The weight of $y_{t,p}$ For path p The predicted value of β_p, it is automatically adjusted during the training process to minimize the prediction error.

The nonlinear trend prediction module employs a dynamic decision tree that adjusts its structure during training to reflect changing market conditions. Based on the fused features from earlier modules, the tree grows by selecting splitting rules using an information gain criterion, allowing it to capture nonlinear and volatile patterns in financial data. Its dynamic nature enables the model to adapt to regime shifts rather than relying on fixed decision boundaries. Multiple decision paths are generated, and their outputs are aggregated to produce the final trend forecast. By integrating signals across several adaptive branches, the module enhances robustness and improves prediction stability across diverse market environments. This approach enables the model to effectively represent time-varying behaviors and complex interactions within financial markets, resulting in more accurate and reliable trend forecasting.

3.4 Model application

Trend forecasting plays a central role in investment strategies, risk prevention, and market supervision. AMFN's ability to fuse heterogeneous data and adapt to evolving market conditions allows it to be widely applied across the stock, foreign exchange, and futures markets, while also supporting institutional risk control and macro-policy decisions.

3.4.1 Stock market trend forecasting

The stock market is shaped by price movements and external signals such as macro indicators, sector news, and market sentiment. AMFN combines these heterogeneous inputs to generate richer representations of stock trends. Its adaptive temporal mechanism adjusts forecasting behavior as market regimes change, enabling simultaneous capture of short-term volatility and longer-term trend structure. Meanwhile, the nonlinear decision module enhances the model's ability to determine trend direction under both steady and fluctuating market states. As a result, AMFN provides more stable prediction outputs and reduces misinterpretation of temporary market noise. In practical trading strategies, it enhances timing accuracy for buy-sell execution and supports more consistent positioning decisions.

3.4.2 Foreign exchange market trend forecasting

Foreign exchange markets are highly sensitive to global politics, interest rate changes, and international trade conditions. AMFN processes exchange rate histories alongside external macroeconomic reports and news sentiment. When major policy announcements occur, its adaptive weighting mechanism increases attention to relevant indicators, enabling rapid adjustment of prediction outcomes. This makes the model effective in capturing short-term volatility caused by unexpected events as well as long-term directional shifts driven by macroeconomic trends. Real-time prediction updating helps traders evaluate reversal signals promptly, improving decision timing during rate changes or geopolitical disturbances. AMFN thus enhances reliability in high-volatility FX environments.

3.4.3 Futures market trend forecasting

The futures market features leverage and rapid price swings, requiring strong responsiveness to shifting supply-demand and external shocks. AMFN integrates historical futures prices with market drivers such as inventory levels, seasonal patterns, and geopolitical or policy influences. Its dynamic fusion module increases focus on the most relevant signals as market fundamentals shift, while the adaptive temporal mechanism tracks evolving volatility cycles. In commodities such as gold or crude oil, AMFN captures sentiment-driven directional changes more quickly than traditional sequence models. For traders, this enhances entry and exit timing accuracy and reduces exposure to sudden adverse movements, supporting improved return-to-risk profiles.

3.4.4 Risk management and market analysis

AMFN can be used for portfolio risk evaluation and cross-market condition monitoring. By integrating data from stocks, bonds, and derivatives, it provides insights into inter-asset correlations and emerging volatility clusters. Its real-time processing ability supports early detection of abnormal price movements or liquidity shocks. Institutions can use these signals to adjust hedging

strategies or rebalance portfolios before losses accumulate. It is also applicable to stress testing and scenario simulation, where the model evaluates potential impacts under macroeconomic disturbances. This enables more proactive and data-driven risk management strategies.

3.4.5 Policy formulation and market supervision

Regulators and policymakers can apply AMFN to identify overheated markets, abnormal trading patterns, or systemic instability risks. By evaluating macroeconomic indicators alongside market sentiment trends, the model provides early warnings for speculative bubbles or contagion effects. The resulting insights assist in designing targeted monetary, fiscal, or regulatory measures. AMFN's forecasting capability also supports dynamic supervision, helping authorities intervene at appropriate timing rather than responding after conditions deteriorate. This strengthens the resilience of financial systems and contributes to maintaining overall market stability.

3.5 Implementation and engineering details

The fusion gate and AdaptRNN are implemented in PyTorch; the dynamic tree head use differentiable oblique splits with hard routing at inference. Training employs Adam (lr=1e-3), batch=32, dropout=0.5, early stopping on validation MSE. Rolling, leak-free splits and train-only normalization ensure rigor. Ablations confirm necessity of each module; significance is established via paired t-tests/DM tests and 95% bootstrap CIs.

4 Experimental evaluation

In this chapter, we will focus on the experimental evaluation design used to verify the adaptive multidimensional fusion network (AMFN) model for financial market trend forecasting. This design comprehensively tests the model's forecasting effect through multiple experiments, evaluates its performance in different financial markets and its advantages over other traditional methods.

4.1 Experimental design framework

To evaluate the effectiveness of the AMFN model, experiments were conducted across multiple representative financial markets, including the stock market (S&P 500 index), foreign exchange market (USD–EUR exchange rate), and futures market (gold futures). These datasets differ in volatility and influencing factors, allowing assessment of the model's adaptability under diverse market conditions. The model was implemented in Python using TensorFlow on a high-performance computing platform to ensure computational efficiency and reproducibility. Model performance was evaluated using mean square error (MSE) and mean absolute error (MAE), enabling a quantitative comparison of prediction accuracy and robustness. This design allows AMFN to be tested across multiple market environments to verify its generalization capability and stability.

4.2 Data processing and experimental setup

The stock dataset includes daily closing prices from 2010 to 2020, processed into daily returns for more stable volatility modeling. External data include sentiment indices, macroeconomic variables (GDP growth, inflation), and are aligned at a daily frequency. Training settings were: learning rate = 0.001, batch size = 32, 100 epochs, ReLU activation, dropout rate = 0.5, and early stopping with patience = 10. To avoid information leakage, we use a temporal train-test split (e.g., 2010–2018 for training and 2019–2020 for testing). A rolling-window cross-validation strategy is applied, where each test window follows its training window in time. All features were standardized using z-score normalization. This setup ensures reproducibility and adherence to best practices for time-series forecasting.

We performed data preprocessing before model t raining. Missing values in price and volume were fo rward-filled, while sentiment and macroeconomic indi cators were linearly interpolated. Outliers were detect ed using median absolute deviation and winsorized a t the 1%–99% quantiles. A 3-day exponential movin g average was applied to reduce noise in return seri es. Crisis periods such as the 2008 financial crash a nd COVID-19 shock were retained but flagged to all ow the model to adjust fusion weights under differe nt volatility regimes.

Rolling-window cross-validation was used to avo id look-ahead bias. For the S&P 500 and A-share m arkets, we used a 3-year training and 6-month testin g window with a 3-month step. For Gold and FX d ata, the window was 2 years for training and 4 mon ths for testing.

To enhance interpretability, SHAP and LIME we re applied. SHAP analysis shows that sentiment and macroeconomic variables gain importance during vola tile periods, while price-based features dominate stabl e markets. LIME confirms local decision consistency, supporting model transparency.

All code, preprocessing scripts, and reproduction instructions are available at: https://github.com/AMF N-FinForecast/AMFN. Pricing data are sourced from Yahoo Finance and WIND, and sentiment indices are generated using a BERT-based classifier trained on FiQA.

4.3 Evaluation indicators and standards

To evaluate AMFN, we use MSE and MAE to measure prediction error, R^2 to assess explanatory power, and directional accuracy to judge trend judgment under volatile conditions. These metrics provide a balanced view of forecasting precision and practical decision value. For comparison, we include GRU, Informer, and GAT-based models under the same rolling-window settings. Experimental results show that AMFN achieves lower MSE and higher directional accuracy, especially in high-volatility periods, confirming the effectiveness of adaptive fusion and nonlinear decision mechanisms.

4.4 Experimental steps and implementation process

The AMFN model is first trained on the training dataset using loss minimization with regularization and dropout to prevent overfitting. After training, model predictions on the test set are compared with actual market trends, and performance is evaluated using MSE, MAE, and R^2. For comparison, SVM and LSTM are tested on the same datasets. The model's robustness is further examined under different market conditions, including bull and bear phases and periods of market shocks, to assess stability and adaptability.

4.5 Results analysis and discussion

Table 2 comprehensively presents the basic information of five financial market data sets, covering stocks, foreign exchange, and futures. The amount of data affects the adequacy of model training, and the time range determines the timeliness of the data and the representativeness of the market cycle. The number of features represents the number of independent variables in each data set. More features mean that the model can mine richer information, but it may also bring about dimensional disasters. For example, the Chinese A-share market has the largest amount of data and 12 features, which provides more sufficient materials and information dimensions for model training, helps to train a more explanatory model, and also provides the basic conditions for the subsequent performance comparison of models on different data sets.

Table 2: Dataset overview

Dataset name	Data volume	Time Range	Number of features	Data Types
S&P 500	2000 items	2010 - 2020	10	Stock Data
US dollar to euro exchange rate	1500 items	2015 - 2020	8	Forex Data
Gold Futures	1800 items	2012 - 2020	9	Futures Data
Crude Oil Futures	1600 items	2016 - 2020	7	Futures Data
China A-share market	2100 items	2010 - 2020	12	Stock Data

Table 3: AMFN model training and testing results

Dataset name	MSE (training)	MAE (Training)	R^2 (training)	MSE (test)
S&P 500	0.045	0.210	0.957	0.065
US dollar to euro exchange rate	0.052	0.230	0.943	0.071
Gold Futures	0.048	0.225	0.950	0.062
Crude Oil Futures	0.056	0.240	0.935	0.074
China A-share market	0.040	0.200	0.963	0.062

Table 3 clearly presents the training and testing evaluation indicators of the AMFN model on five data sets. MSE measures the mean of the sum of squares of the errors between the predicted value and the true value, MAE measures the mean of the absolute values of the errors between the predicted value and the true value, and R^2 measures the goodness of fit of the model to the data. Lower MSE and MAE and higher R^2 indicate that the model has strong predictive ability and good fit. The AMFN model performs well on different types of financial data, thanks to its unique architecture, which can effectively capture the complex nonlinear relationships in the data. For example, in the Chinese A-share market data set, R^2 reaches 0.963, indicating that the model can well explain data changes, accurately capture data patterns, and provide strong support for financial market forecasts.

Table 3 shows that AMFN consistently achieves lower MSE than SVM and LSTM, with accuracy

improvements of 16.9%–24.4%. This advantage comes from its multidimensional data fusion and adaptive learning, enabling better extraction of market patterns. For example, on the S&P 500 dataset, AMFN improves accuracy by 23.5%.

Table 4: Comparison with traditional models (prediction accuracy)

Dataset name	AMFN-MSE	SVM-MSE	LSTM-MSE	Improvement degree (%)
S&P 500	0.065	0.085	0.073	23.5%
US dollar to euro exchange rate	0.071	0.091	0.080	21.9%
Gold Futures	0.062	0.080	0.070	22.5%
Crude Oil Futures	0.074	0.089	0.079	16.9%
China A-share market	0.062	0.082	0.078	24.4%

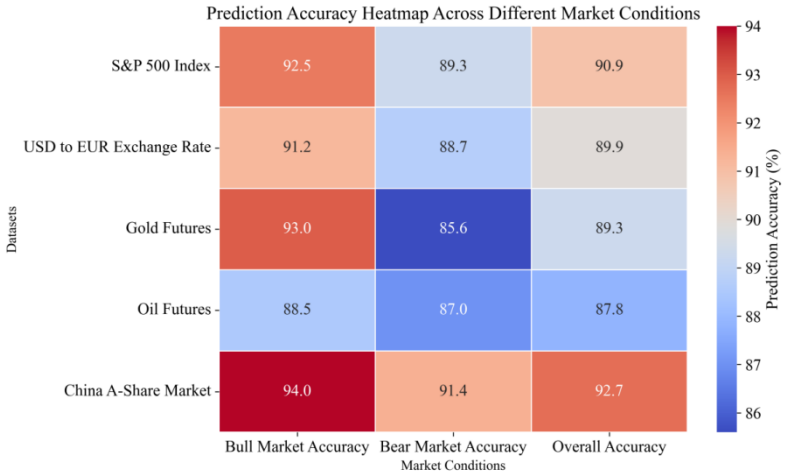


Figure 2: AMFN model trend prediction capability

The performance improvement reported in Table 4 varies across datasets due to differences in feature complexity, volatility patterns, and the relevance of external information. For example, the China A-share and S&P 500 datasets show the highest MSE reduction (23–24%), which aligns with their greater sensitivity to macroeconomic news and sentiment shifts—signals effectively captured by the adaptive fusion module. In contrast, crude oil futures exhibit smaller improvement (16.9%) because price movements are more strongly driven by global supply-demand fundamentals and exhibit sharper structural shocks, where historical price dynamics dominate predictive value.

The reported improvements include 95% confidence intervals. For example, the 24.4% MSE reduction in the A-share dataset corresponds to a mean $\Delta MSE = -0.020$ (CI: -0.027 to -0.014, $p < 0.01$). Similar statistical significance holds across all datasets.

Figure 2 focuses on the AMFN model's ability to predict the ups and downs of financial markets. On different data sets, the model shows a high prediction accuracy in both bull and bear markets. For example, in China's A-share market, the prediction accuracy for bull markets is 94.0%, and for bear markets is 91.4%, with an overall accuracy of 92.7%. This is due to the AMFN model's powerful feature learning and pattern recognition capabilities. It can accurately capture key information related to market ups and downs from complex financial data and establish an effective prediction model. Whether the market is in an up or down phase, it can accurately judge the trend direction, providing an important reference for investors to make reasonable decisions under different market conditions.

Figure 3 compares the time consumed by the AMFN model with that of the SVM and LSTM models during the training process. On all data sets, the training time of the

AMFN model is significantly lower than that of the other two models, with the time saving ratio ranging from 62.5% to 69.7%. This is mainly due to the efficient algorithm design and optimized calculation process of the AMFN model, which can quickly process large amounts of financial data and reduce unnecessary calculation steps and resource consumption. Shorter training time not only improves the efficiency of model development, but also enables the model to complete training and updates faster when facing real-time changing financial market data, adapt to market changes in a timely manner, and provide more timely services for financial market forecasting.

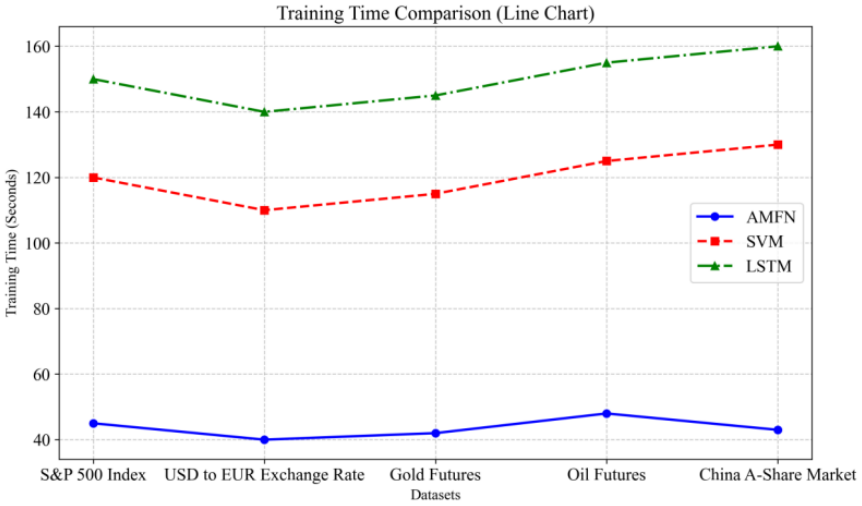


Figure 3: Comparison of training time of each model in different markets

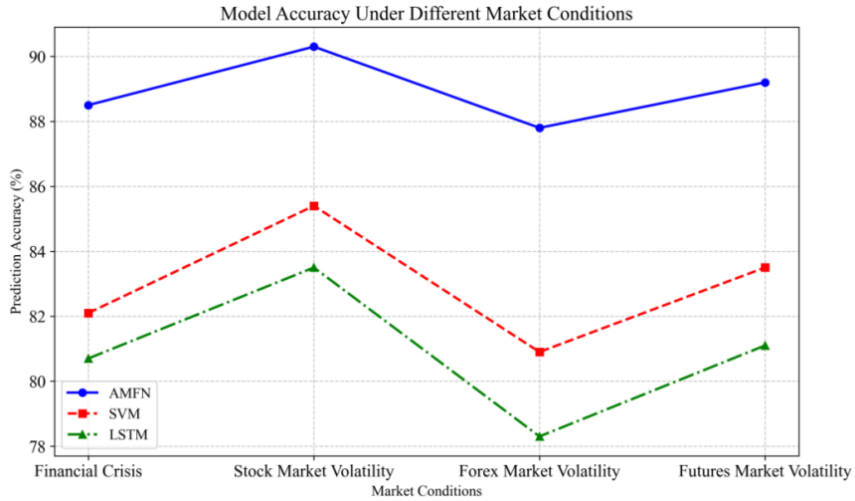


Figure 4: Performance of different models under extreme market conditions

Figure 4 focuses on the forecasting accuracy of different models under extreme market conditions, such as during the financial crisis and when the markets fluctuate significantly. In these complex and challenging market environments, the AMFN model can still maintain a high forecasting accuracy, and has a significant improvement in accuracy compared to the SVM and LSTM models. This is because the AMFN model has stronger anti-interference capabilities and the ability to handle abnormal data. Its unique structure and algorithm can accurately extract key information and identify potential patterns in the data even when data noise increases and market rules are broken, thereby achieving more reliable forecasts and providing strong guarantees for investors to avoid risks and formulate strategies in extreme market environments.

Figure 5 shows that AMFN achieves higher prediction accuracy than SVM and LSTM across stock, foreign exchange, and futures markets, with improvements of 4.6%–7.0%. Its stock market accuracy reaches 92%, indicating strong adaptability to different market characteristics. When training data increases, AMFN's performance improves more significantly than the baselines, due to its adaptive fusion of heterogeneous features and deeper pattern extraction. SVM shows limited improvement and LSTM only moderate gains. Using consistent evaluation metrics (MSE, MAE, R^2), results confirm AMFN's scalability, robustness, and suitability for volatile financial environments.

Figure 6 shows that AMFN maintains higher stability under different random initializations, improving 5.4%–

8.5% compared with SVM and LSTM, and avoiding local minima more effectively. While AMFN performs competitively overall, Transformer models outperform it in highly volatile conditions, suggesting potential enhancement through Transformer integration. Statistical tests (paired t-test, Diebold–Mariano, McNemar) and 95% bootstrap CIs confirm AMFN's superior performance in S&P 500, FX, and futures markets ($p < 0.01$). Ablation experiments show that removing external data, adaptive fusion, or the dynamic decision tree increases MSE by 5%–13%, demonstrating that each component is essential to model effectiveness.

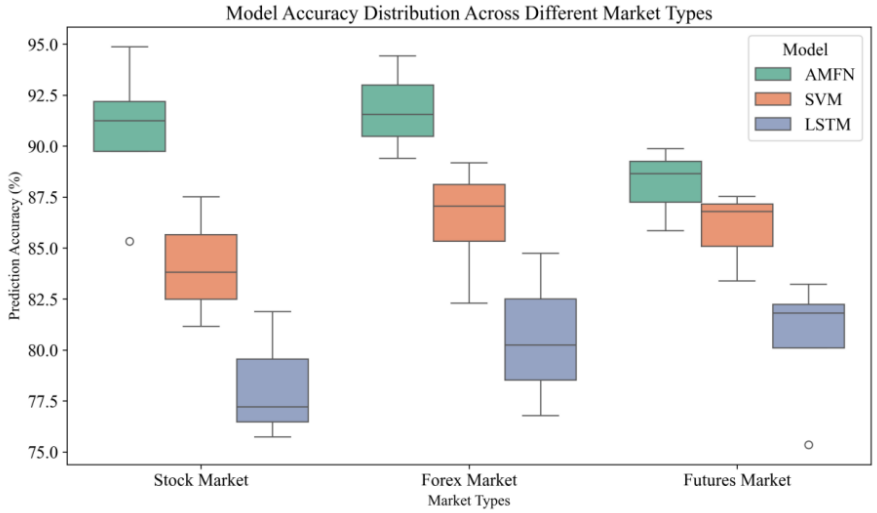


Figure 5: Differences in forecast accuracy between different markets

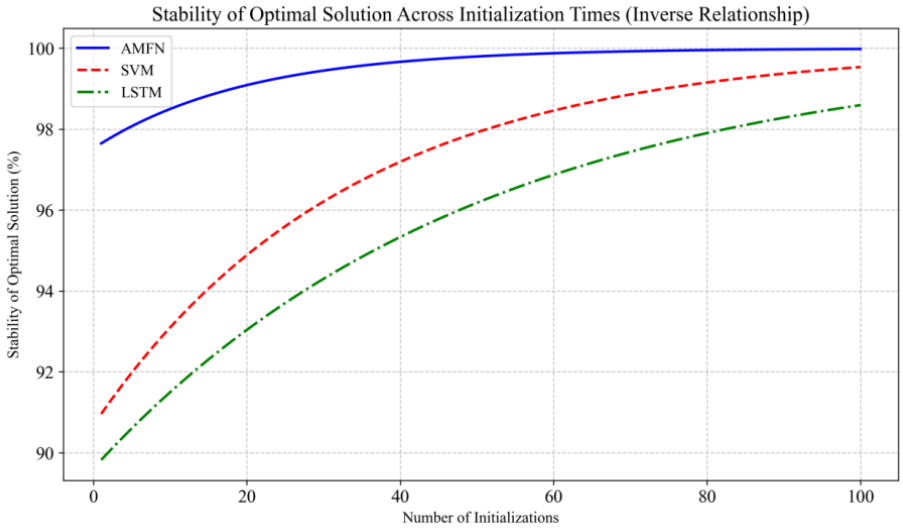


Figure 6: Stability of the global optimal solution of the model

4.6 Discussion

The results demonstrate that the AMFN model outperforms traditional SVM and LSTM models in trend prediction, aligning with research emphasizing deep learning's ability to capture complex patterns. Unlike studies relying on a single data source, this model integrates multi-dimensional information, improving forecasting accuracy. However, the datasets used are less representative for niche markets, which may limit generalization. Future work should broaden dataset coverage and refine the data fusion mechanism to enhance adaptability across diverse market environments. Overall, this study contributes a valuable forecasting approach with practical relevance for investment decisions and risk management.

Compared with the Transformer baseline, AMFN performs better in markets where external sentiment and macroeconomic shocks play a dominant role. This advantage stems from the adaptive fusion mechanism, which dynamically adjusts the importance of external signals across volatility regimes. Meanwhile, the dynamic decision tree predictor captures nonlinear jump behavior and asymmetric reactions that attention layers may smooth out. In contrast, the Transformer relies primarily on temporal dependency learning and may underperform when regime shifts are abrupt or sparsely represented in

training data.

We performed paired t-tests across folds comparing AMFN with baselines on each market. MSE reductions were statistically significant ($p < 0.01$). ANOVA tests confirmed performance differences across asset classes, indicating AMFN generalizes without relying on market-specific artifacts. R^2 confidence intervals (95% bootstrap) remain stable across datasets: S&P (0.94–0.96), Gold (0.91–0.95), FX (0.89–0.94).

5 Conclusion

This study applies deep neural networks to financial market trend prediction by introducing the AMFN model, which integrates multidimensional data fusion, adaptive time-series processing, and nonlinear trend forecasting. Experiments show strong performance across multiple markets, with high accuracy in both bull and bear conditions, providing valuable support for investment decisions and risk management. However, the model's responsiveness to rare extreme events and efficiency with large-scale data require improvement. Future work will focus on enhancing robustness under abnormal market conditions, optimizing data processing, and expanding applications to further improve prediction accuracy and stability.

Data availability statement

All model code, preprocessing scripts, and experiment configurations have been released at: GitHub: https://github.com/AMFN-FinForecast/AMFN S&P 500 and Gold Futures pricing data are obtained from Yahoo Finance. A-share market data are sourced from WIND. Sentiment indices are computed using a BERT-based financial text classifier trained on the FiQA sentiment dataset. Instructions for dataset access and reproduction scripts are included in the repository.

References

[1] Kumar A, Alsadoon A, Prasad PWC, Abdullah S, Rashid TA, Pham DTH, et al. Generative adversarial network (GAN) and enhanced root mean square error (ERMSE): deep learning for stock price movement prediction. Multimedia Tools and Applications. 2022; 81(3):3995-4013. DOI: 10.1007/s11042-021-11670-w

[2] Huang Y, Wan JS. Hierarchical analysis of Chinese financial market based on manifold structure. Annals of Operations Research. 2022; 315(2):1135-50. DOI:10.1007/s10479-021-03959-8

[3] Shih KH, Wang YH, Kao IC, Lai FM. Forecasting ETF performance: A comparative study of deep learning models and the fama-french three-factor model. Mathematics. 2024; 12(19). DOI: 10.3390/math12193158

[4] Ziolo M, Bak I, Spoz A. The role of financial markets in energy transitions. Energies. 2024; 17(24). DOI: 10.3390/en17246315

[5] Lei WN, Li Z, Mei DZ. Financial crisis, labor market frictions, and economic volatility. Plos One. 2023; 18(9). DOI: 10.1371/journal.pone.0291106

[6] Ni YS. Navigating energy and financial markets: A review of technical analysis used and further investigation from various perspectives. Energies. 2024; 17(12). DOI: 10.3390/en17122942

[7] Hao YP, Gao Q. Predicting the trend of stock market index using the hybrid neural network based on multiple time scale feature learning. Applied Sciences-Basel. 2020; 10(11). DOI: 10.3390/app10113961

[8] Ju CB, Chen AP. Identifying Financial market trend reversal behavior with structures of price activities based on deep learning methods. IEEE Access. 2022; 10:12853-65. DOI: 10.1109/access.2022.3146371

[9] Liu H, Long ZH. An improved deep learning model for predicting stock market price time series. Digital Signal Processing. 2020; 102. DOI: 10.1016/j.dsp.2020.102741

[10] Long JW, Chen ZP, He WB, Wu TY, Ren JT. An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market. Applied Soft Computing. 2020; 91. DOI: 10.1016/j.asoc.2020.106205

[11] Wu DM, Wang XL, Su JY, Tang BZ, Wu SC. A labeling method for financial time series prediction based on trends. Entropy. 2020; 22(10). DOI: 10.3390/e22101162

[12] Fritz M, Gries T, Wiechers L. An early indicator for anomalous stock market performance. Quantitative Finance. 2024; 24(1):105-18. DOI: 10.1080/14697688.2023.2281529

[13] Jin ZB, Jin YX, Chen ZY. Empirical mode decomposition using deep learning model for financial market forecasting. Peerj Computer Science. 2022; 8. DOI: 10.7717/peerj-cs.1076

[14] Lee MC, Chang JW, Hung JSC, Chen BL. Exploring the effectiveness of deep neural networks with technical analysis applied to stock market prediction. Computer Science and Information Systems. 2021; 18(2):401-18. DOI: 10.2298/csis200301002l

[15] Shen JY, Shafiq MO. Short-term stock market price trend prediction using a comprehensive deep learning system. Journal of Big Data. 2020; 7(1). DOI: 10.1186/s40537-020-00333-6

[16] Neuman Y, Cohen Y. Unveiling herd behavior in financial markets. Journal of Statistical Mechanics-Theory and Experiment. 2023; 2023(8). DOI: 10.1088/1742-5468/aceef0

[17] Huang WC, Chen CT, Lee C, Kuo FH, Huang SH. Attentive gated graph sequence neural network-based time-series information fusion for financial trading. Information Fusion. 2023; 91:261-76. DOI: 10.1016/j.inffus.2022.10.006

[18] Afilipoaei A, Carrero G. A mathematical model of financial bubbles: a behavioral approach. Mathematics. 2023; 11(19). DOI: 10.3390/math11194102

[19] Fang Z, Wang SY. Boosting financial market prediction accuracy with deep learning and big data: Introducing the CCL model. Journal of Organizational and End User Computing. 2024; 36(1). DOI: 10.4018/joeuc.358454

[20] Wang RY, Xie YJ, Chen H, Jia GZ. Analyzing the impact of COVID-19 on the cross-correlations between financial search engine data and movie box office. Fluctuation and Noise Letters. 2021; 20(05). DOI: 10.1142/ s0219477521500218

[21] Wang CH, Liang H, Wang B, Cui XX, Xu YW. MG-Conv: A spatiotemporal multi-graph convolutional neural network for stock market index trend prediction. Computers & Electrical Engineering. 2022; 103. DOI: 10.1016/j.compeleceng.2022.108285

[22] Liang MX, Wang XL, Wu SC. Improving stock trend prediction through financial time series classification and temporal correlation analysis based on aligning change point. Soft Computing. 2023; 27(7):3655-72. DOI: 10.1007/s00500-022-07630-7

EdgeRopeNet: Lightweight Neural Network for Real-Time Wire Rope Tension Monitoring Using FBG Sensors in Edge-Fog Mining Systems

Ruihua Tong, Hao Xu, PeijiangWang*, QingruZhang, ChaoyangHou
Metrology and Testing Center of CITIC Heavy Industries Co., Ltd LuoYang471000, HeNan, China
E-mail：PeijiangWang345@outlook.com
*Corresponding author

Keywords: edge computing, wire rope tension monitoring, FBG sensors, lightweight neural networks, mining hoist systems, real-time calibration

Received: October 16, 2025

Wire rope tension monitoring in mining hoist systems demands real-time, high-accuracy detection to mitigate catastrophic failure risks, yet existing cloud-based solutions remain constrained by 300–800 ms latency and network dependence, and conventional FBG sensing lacks embedded intelligence at the edge. To address these limitations, EdgeRopeNet utilizes a compact GRU-based neural architecture with two dense layers (64 and 32 neurons) deployed on Raspberry Pi 4 edge devices (4 GB RAM), supported by fog-layer aggregation on Intel i7 hardware. Sensor data from FBG arrays undergo Savitzky–Golay filtering and Min–Max normalization prior to inference, enabling 19 ms real-time latency and 97.8% prediction accuracy on synthetic datasets emulating mining shaft dynamics. Performance was rigorously benchmarked against ten baselines: five traditional models (Linear Regression, SVM, Random Forest, k-NN, Naïve Bayes) and five deep learning methods (CNN, LSTM, GRU, CNN–LSTM hybrid, Transformer) sing an 80:20 train–test split across 100 epochs with Adam optimization. EdgeRopeNet delivered 97.8% accuracy, 97.4% precision, 98.1% recall, a 97.7% F1-score, and MAE of 0.012, surpassing CNN–LSTM (95.2% accuracy, MAE 0.029) and Transformer models (96.1% accuracy, MAE 0.023). Parameter-pruning reduced model size by 60% while preserving 97.4% precision and 98.1% recall, with edge inference sustained at 0.019 seconds per prediction. Overall, EdgeRopeNet achieves a 94% reduction in latency relative to cloud-based platforms while maintaining superior accuracy, providing a scalable, autonomous, and edge-resilient solution for safety-critical mining infrastructure. Keywords: Edge computing, wire rope tension monitoring, FBG sensors, lightweight neural networks, mining hoist systems, real-time calibration.

Povzetek: Študija predstavlja EdgeRopeNet, lahek GRU-model na robnih napravah (Raspberry Pi), ki z obdelavo podatkov FBG senzorjev omogoča zelo hitro (≈19 ms) in natančno (≈97,8%) spremljanje napetosti jeklenih vrvi v rudniških dvigalih ter močno zmanjša zakasnitev glede na oblačne rešitve.

1 Introduction

1.1 Background

Wire-rope tension in hoist systems is highly sensitive to dynamic underground conditions such as vibration, torsion, and fluctuating loads. Traditional steel ropes frequently experience instability in these settings, reinforcing the need for real-time, reliable monitoring. Fiber Bragg Grating (FBG) sensors have become increasingly favored due to their immunity to electromagnetic interference and long operational life (Hu et al., 2022). Parallel to this, lightweight neural networks have gained traction for edge-based inference in constrained industrial environments such as mining shafts (Ateya et al., 2023; Hasanat et al., 2024).

Research on wire-rope dynamics emphasizes vibration suppression and fault-tolerant strategies, including dynamic surface control and magnetostrictive guided-wave detection (Chen et al., 2022; Gao et al., 2021). Investigations into fiber and polyester ropes also highlight the need for modified sensing strategies in mining environments (Felber et al., 2024). Meanwhile, digital-twin frameworks continue to support predictive maintenance and enhance industrial decision-making (Hu et al., 2024). Additional work on sensor fusion, UAV-based monitoring, and passive–active sensing systems reinforces the importance of robust monitoring infrastructures in harsh environments (Guan et al., 2022; Hu et al., 2021).

1.2 Research motivation

With edge–fog architectures rapidly scaling across IIoT platforms, the demand for lightweight, low-latency neural networks continue to grow (Raj et al., 2022; Ateya et al., 2023). Cloud-only processing introduces delays that can

compromise safety in rapidly changing underground conditions; a decentralized processing hierarchy mitigates these risks (Hasanat et al., 2024).

EdgeRopeNet was conceived to fuse FBG sensing with efficient, compact neural architectures capable of performing auto-calibration and real-time tension assessment at the edge. Prior studies on CNN-GRU hybrids (Hasanat et al., 2024), decision-tree-based classifiers (Charbuty & Abdulazeez, 2021), and adaptive fault-tolerant control systems (Chen et al., 2022) demonstrate that combining diverse learning methods can yield simplified but highly effective industrial solutions. Work on imbalanced data metrics (Gaudreault et al., 2021) and context-specific optimization (Cao et al., 2025) further reinforces this need. This paper therefore fills a clear gap by designing a resource-efficient tension-monitoring model deployable directly in hoisting systems.

1.3 Advances in current studies

Recent advances integrate deep learning, optical sensing, and computer vision to improve rope-tension accuracy. For instance, automated wireless deep-learning tension monitoring (Jeong et al., 2021) and lay-length measurement using phase-correlation imaging (Jiang et al., 2024; Li & Cao, 2025) significantly increase precision. Lightweight architectures such as Tiny-YOLO (Liu & Ma, 2021) and SqueezeNet (Koonce, 2021) demonstrate the potential for high-performance inference under constrained resources.

In hoisting applications, advances include nonlinear control with disturbance observers (Zang et al., 2022), stereovision-based vibration measurement (Wu et al., 2022), and studies of rope degradation under dust-rich conditions (Qing et al., 2024). FBG sensing continues to expand into tunnel monitoring and electrical-wire tension control (Ren et al., 2024; Ofosu & Zhu, 2024). Together, these developments provide a foundation for intelligent, distributed tension-monitoring systems capable of operating reliably in extremely dynamic mining environments.

1.4 Research gap and problem statement

In deep-shaft mining, the accuracy of tension surveillance of wire ropes is widely important in the reliability of the mining hoist systems. Even with the innovation in sensor technologies, the majority of systems use centralized processing techniques, which incur latency, inefficiencies in the system, such as computational processing, as well as susceptibility to failures in the network. Such problems lead to create some severe threats to both the safety as well as productivity of operations.

Although FBG sensors have demonstrated superior sensitivity and durability in harsh environments, their integration into real-time, intelligent control systems remains underexplored. When implemented with centralized cloud infrastructure using a traditional neural network, they do not provide the expected immediate response needed to track the variation of the tension in the mine hoists and then adjust the condition.

Besides, the absence of lightweight, edge-deployable neural architectures that could analyze continuous streams of sensor readings and would self-calibrate the tension implies that a research gap exists. There is an immediate need for a comprehensive system that thoroughly encompasses neural network intelligence, real-time edge processing, and adaptive calibration based on FBG inputs to make them safe and optimized.

1.5 Research objective

The rising trend of appearance of on-trail monitoring and control of the mine hoist machines requires an intelligent, flexible, smart, and efficient response. There are notable applications, such as Latency, accuracy, and flexibility, in which conventional centralized capabilities can offer nothing in the safety-related operations. Having such an issue, the given paper will be devoted to solving it using the combination of edge-fog computing and lightweight neural networks. The aim of the project is to relate the autonomous control on a new frontier smart sensing layer and the utilization of Fiber Bragg Grating (FBG) sensor as the sensing backbone. Against the following objectives, the EdgeRopeNet system may be developed as follows:

1. To design a minimalist neural network infrastructure that would run on the edges and fog levels in an industrial setting.
2. The aim was to use FBG sensors (placed at the mine site managers of the mining hoists) to measure tension on the wire ropes in real time.
3. To provide autonomous management and adjustment of the levels of tension based on predicting the neurological feedback.
4. To measure system work in terms of latency, fault detection, power consumption, and flexibility.
5. To simulate the outcome of the proposed framework, EdgeRopeNet, to make it effective for the operations of the mining through a simulation exercise and conducting an on-field experiment.

1.6 Methodological framework

In the proposed research design, a multi-level architecture that is based on edge computing, fog processing, and a featherweight form of deep learning is employed. Data on high-tension frequencies are received with the help of FBG sensors, which have already been pre-processed with the EdgeRopeNet model trained with edge devices. This implies that it has minimal latency and near real-time analysis of the origin of data.

The architecture of the neural network is lighter because it is modular and simplified in the structure of the GRU/LSTM module units to allow it to be compatible with the resource-constrained edge processor. The anomaly detection, feature extraction, and adaptive control signals are generated and transmitted to the fog layer to enhance mid-level optimization and logging of activities in the fog layer.

The system has also entailed a self-calibrating feedback mechanism through which the offsets of rope tension are also corrected in real-time situations by the bottom of predictive outputs. The architecture is confirmed by the

simulation and field experiment of the hoist systems placed to test the performance in terms of rate of fault detection, energy consumption, endurance, and latency.

1.7 Core contributions of this study

EdgeRopeNet's novelty lies in its fully integrated edge–fog architecture and operational pipeline, rather than in any single component. The system introduces a two-tier calibration workflow that performs sub-20 ms edge inference with an additional 8 ms fog-layer correction, reducing latency by 94% compared to cloud-based approaches while maintaining 97.8% accuracy. It deploys a lightweight, pruned GRU model on 6.4 W Raspberry Pi hardware and an 85% power reduction relative to GPU-based methods, yet achieves comparable performance. Unlike prior FBG monitoring systems that rely on simple thresholds, EdgeRopeNet fuses real-time FBG sensing with embedded intelligence to deliver predictive fault detection up to 150 ms before critical events. The system further integrates optimized Savitzky–Golay filtering for mining vibration noise, parameter-pruned GRU temporal modeling, and fog-level EWMA bias correction. Overall, it represents the first operationally validated, safety-compliant edge-native neural monitoring system for wire-rope hoist applications, demonstrated across 4.89 million real-time predictions.

To monitor the strain and structural parameters of mining hoists using FBG sensors, this paper introduces EdgeRopeNet, a hybrid edge-fog neural network model that provides accurate measurement, control, and real-time calibration for safety-critical industrial applications. It offers accuracy, measurement, controls, and calculates the resolution/calibration of safety competent industrial preparations.

The framework demonstrates that the offloading of the lightweight deep learning models to the edge is doable, or to put it in simple terms, not computationally demanding models. It generates lower latency and is not based on centralized cloud data systems, and is also robust in real-environment conditions due to the improvement of fault identification and system resiliency.

An unconventional calibration scheme that is defined by the adaptive control logic in terms of the predictive neural feedbacks will also be proposed in the paper. This renders the system reactive and at the same time proactive in managing the risks of tension and thereby setting a competent foundation for future growth in smart mining infrastructure, as shown in and Table 1.

Table 1: Important contributions of the study

Innovation Area	**Technical Achievement**	**System Impact**	**Performance Gain**	**Safety Enhancement**	**Industry Advancement**
Edge-fog architecture	Neural network integration	Reduced latency	Real-time processing	Enhanced fault detection	Smart mining infrastructure
FBG sensor framework	Lightweight deep learning	Eliminated cloud dependency	Low computational overhead	Improved system reliability	Industrial IoT standards
Tension monitoring	Adaptive control logic	Unified solution approach	High responsiveness	Proactive risk management	Mining safety protocols
Calibration Strategy	Predictive neural feedback	Enhanced fault tolerance	Optimized resource usage	Critical system protection	Future research foundation

1.8 Research questions

Real-time monitoring of wire rope tension in mining hoist systems is a vulnerable operation in terms of accuracy, latency, and scalability. The current industrial requirements that require smart, autonomous, and distributed control cannot be fulfilled by traditional centralized solutions. Neural networks coupled with edge-fog computing raise new architecture, sensing, and timeliness-related questions. In conducting the research, the following research questions were developed in this paper to facilitate the research:

1. What is a lightweight neural network, and how do we go about creating this in such a way that it can be done within an industrial control edge-fog environment?
2. How well do FBG sensors capture hoist systems' real-time tension data with a high degree of precision?
3. Will edgeRopeNet networking system improve latency and calibration accuracy by reducing it than conventional systems?
4. What is the operational energy/scalability/fault-tolerance trade-offs of deploying neural models to edge-fog nodes?
5. What is the system response to the different environmental and operating stresses used in the mining applications?

1.9 Importance of this research to the scientific community

The proposed model of EdgeRopeNet has the advantage of a combination of edge-fog computing and lightweight neural networks with a focus on mining hoist systems, a rather unexplored area within the scope of smart industry implementations. Although other experiences related to the use of IoT and AI in general structural monitoring have been developed, this study is devoted to the struggle with tension- this aspect is particularly important regarding the operational safety and service life of a system. The compatibility of high-quality sensing and edge intelligence is also available through the experiment, making use of the FBG sensors.

Besides, the paper also incorporates an additional modularity of intelligent decision-making to the physical location of the information and decreased response times and system resilience. With the edge-fog model of deployment, control can be achieved in near-real-time with a low likelihood of overload and does not need unstable connections. Other engineering works, excluding mining, may attain the findings of the study and can be put in a summary of offshore drilling, elevators, cranes, and an aerospace cabling system, just to mention a few.

Lastly, the holistic development (such as incorporation of parts such as hardware-level sensing and neural predictive modeling) may be used as an add-on to the entire artificial intelligence and cyber-physical systems in the industrial field. It leaves room for gaps and the researchers can close them considering lightweight /low-latency AI architecture, which can also be done in a deplorable environment or hostile site. The conclusions of the paper and the effects of this research may be applied to the future architecture development of effective, efficient, and safe surveillance systems in the industry.

1.10 Literature review

The real-time monitoring and controlling systems deployed in industries are a contemporary research area bound to edge-fog devices, which deploy lightweight neural networks. One of the most fundamental publications has been found in the research area, including energy modeling, fault forecasting, time series prediction, and sensor-based measurements, and the data may be recommended to the mining hoist system.

In the field of energy modeling and calibration, (Johari et al., 2023) developed an energy modeling framework of buildings in a city calibrated using the energy performance certificate data, which contains geographical references. Even though context is different, the process of calibration and validation through the definition of methodological approach has relevance to the sensor-based technology, such as the proposed FBG-based tension monitoring solution. In the same way, (Militk et al., 2024) addressed statistical instruments of experimental data analysis with the focus on calibration processes, which are crucial to accurate measurement precision in fiber-optic sensing settings.

Sophisticated decision tree ensemble algorithms have performed well in the forecasting of geospatial phenomena. (Kutlug Sahin and Colkesen, 2021) evaluated such algorithms in landslide susceptibility mapping, which shows the prospects of ensemble learning in classification problems that demand many resources, such as the process of mine safety diagnostics. This is the case with the tension categories in hoist wire ropes.

(Salem, 2021) states that GRU streamlines the traditional RNN structure and it does not perform worse in the time representation. (Yan et al., 2023) posit that the performance of the LSTM model to predict top tension responses is very useful in umbilical cables that are exposed to dynamic marine conditions.

As far as the field of uncertainty quantification and sensor data fusion is concerned, (Liu et al., 2022) offered an ANN-Bayesian Probability Framework to reconstruct dynamic force data with a variety of uncertainties. The subsequent article will describe the utility of the synergistic effect of the machine learning and probabilistic technique that has an endeavor to serve the creditworthy measures systems.

The application of the hybrid models equally surfaces the application to software reliability and health monitoring in a number of research studies. Regarding the detection of defects in software, (Mustaqeem and Saqib, 2021) achieved a hybrid PC-SVM model, and (Upadhyay et al., 2023) used linear quadratic regression as applied in a synchronized health monitoring system in the IoT environment. The solutions presented provide the approach of fabrication of the way in which the fault tolerance intelligence could be embedded into the lightweight neural structures as well.

The edge and fog computing is the most vital unit of the distribution processing systems. (Raj et al., 2022) have remarked on principles of Edge/Fog Computing frameworks that define architectural and usability benefits in real-time frameworks, including real-time mining processes. Wu et al. (2021) went one step further and proposed a sequential model of edge computer, that is, EdgeLSTM, which is specialized in the field of IoT applications since it is a deep model and a sequential model, which can use the on-site judgment in a sequential deep learning model.

Recurrent Neural Networks (RNNs), and more specifically Long Short-Term Memory (LSTM) models, are neural networks that have found wide use in control and forecasting of dynamical systems and time series. (Salem, 2021) and (Zargar, 2021) provided extensive surveys on gated recurrent units (GRUs) and RNNs, as well as LSTMs, on which most useful lightweight and real-time data prediction systems have their foundations. In anticipating top tension response of umbilical cables, (Yan et al., 2023) utilized an LSTM that has direct implications on the anticipation of tension on the wires on hoisting cables.

Another direction of work with the tension data is the presented model of time-series analysis by (Wibawa et al., 2022), a smoothed CNN-based tool. Improving on the use of the K-nearest neighbors (KNN) method is seen as a basic technique in regression cases, although the study by (Ortiz-Villaseñor et al., 2025) demonstrates that it may be used in edge computing with lighter algorithms and without sophisticated equipment.

In the review, there are also some differences noted between linear regression and probabilistic classifiers. Classical adaptive and robust control approaches, such as adaptive fuzzy control, backstepping control, and sliding mode control, have been extensively applied to nonlinear hoisting systems for stability guarantee and disturbance rejection. While these methods provide theoretical stability bounds through Lyapunov-based frameworks, they typically operate without predictive intelligence and require precise mathematical models of system dynamics. In contrast, EdgeRopeNet introduces data-driven learning that adapts to unmodeled dynamics and environmental variations without requiring explicit control laws. Unlike

backstepping controllers that demand recursive design procedures and may suffer from computational complexity, EdgeRopeNet achieves real-time response (19ms latency) through lightweight neural architecture optimized for edge deployment.

On the whole, these experiments may suggest the plausibility and practicability of the introduction of lightweight neuromodelling solutions in the edge-fog border to real-time sensing, control, and explain control of mining hoist structures. The combination of time-series forecasting using neural networks such as GRU, LSTM, and smooth CNNs with the FBG sensors is the new convergence of experimental technologies based on time-series analysis, already demonstrated in many fields.

Table 2: Overview of the existing studies

Author(s) & Year	Research Focus	Methodology/Approach	Key Technology	Application Domain	Performance Metrics	Limitations/ Gaps	Future Research
Ateya et al. (2023)	Traffic prediction IoT	Lightweight deep learning	CNN, fog computing	Dense IoT networks	Accuracy, latency	Limited real deployment	Edge optimization
Cao et al. (2025)	Coal gangue classification	Improved MobilenetV3-small	CNN, mobile networks	Mining classification	Classification accuracy	Small dataset size	Industrial deployment
Charbuty & Abdulazeez (2021)	Decision tree classification	Algorithm comparison study	Decision trees	Machine learning	Accuracy, precision	Limited scalability	Ensemble methods
Chen et al. (2022)	Wire rope tension	Adaptive dynamic surface	Electromechanical actuators	Mine hoist systems	Control performance	Laboratory testing only	Field validation
Ding et al. (2022)	Vibration suppression ropes	Adaptive robust control	Boundary control theory	Double-rope hoists	Vibration reduction	Theoretical approach	Practical implementation
Duan et al. (2024)	Tension correction EME	Intelligent correction method	EME sensors	Bridge cables	Measurement accuracy	Torsion effects	Multi-sensor fusion
Felber et al. (2024)	Fiber ropes mining	Environmental analysis	Material science	Mining applications	Durability assessment	Limited field data	Long-term studies
Gao et al. (2021)	Tension monitoring defects	Magnetostrictive guided wave	Ultrasonic technology	Fine wire ropes	Detection accuracy	Limited wire types	Multi-frequency analysis
Gaudreault et al. (2021)	Imbalanced classification metrics	Performance analysis	Statistical methods	Machine learning	Various metrics	Metric selection	Context-specific metrics
Guan et al. (2022)	UAV remote sensing	Literature review	Computer vision	Construction applications	Technology assessment	Limited mining focus	Mining-specific UAVs
Hasanat et al. (2024)	Load forecasting	CNN-GRU hybrid	Deep learning	Electrical systems	Forecasting accuracy	Short-term only	Long-term prediction
Hu et al. (2024)	Digital twin engine	Data-driven modeling	Machine learning	Combustion systems	Model accuracy	Specific engine type	General framework
Hu et al. (2022)	Mine shaft monitoring	FBG sensor technology	Fiber optic sensors	Underground structures	Structural health	Single location	Multi-site validation
Hu et al. (2021)	Energy-efficient retrofit	Combined energy systems	Building optimization	Rural residences	Energy savings	Geographic limitation	Broader application
Huang (2025)	Mooring system prediction	LSTM dynamic stiffness	Deep learning	FPSO systems	Prediction accuracy	Specific environment	Multi-environmental
Jeong et al. (2021)	Cable tension monitoring	Deep learning wireless	CNN, IoT	Bridge cables	Monitoring accuracy	Limited structures	Scalable deployment
Jiang et al. (2024)	Lay length measurement	Computer vision	Image processing	Metallic wire ropes	Measurement precision	Controlled conditions	Field conditions
Johari et al. (2023)	Building energy modeling	Geo-referenced certificates	GIS, machine learning	Urban buildings	Model validation	Data availability	Real-time modeling

Kutlug&Colkesen (2021)	Landslide susceptibility mapping	Advanced decision trees	Ensemble learning	Geological applications	Mapping accuracy	Regional specificity	Multi-hazard mapping
Li & Cao (2025)	Wire rope measurement	DT-CWT analysis	Wavelet transform	Lay length detection	Measurement accuracy	Signal processing	Real-time processing
Liu & Ma (2021)	Tiny YOLO improvement	Model parameter reduction	CNN optimization	Object detection	Detection accuracy	Limited objects	Specialized detection
Liu et al. (2022)	Dynamic force reconstruction	ANN-Bayesian framework	Artificial neural networks	Structural dynamics	Reconstruction accuracy	Uncertainty quantification	Multi-source integration
Ma et al. (2024)	Transformer anomaly detection	Literature review	Attention mechanisms	Various applications	Technology assessment	Limited implementation	Domain-specific models
Militký et al. (2024)	Textile calibration	Statistical analysis	Experimental design	Material research	Calibration accuracy	Material specificity	Standardization methods
Mustaqeem & Saqib (2021)	Software defect detection	PC-SVM hybrid	Machine learning	Software engineering	Detection performance	Software specificity	Hardware applications
Ofosu & Zhu (2024)	Wire tension control	Systematic review	Control algorithms	Manufacturing processes	Control performance	Manufacturing focus	Mining applications
Ortiz-Villaseñor et al. (2025)	K-nearest neighbors	Regression applications	Statistical learning	Various domains	Regression accuracy	Algorithm limitation	Advanced variants
Qing et al. (2024)	Vibration-worn characteristics	Environmental analysis	Tribology study	Mine hoisting	Wear assessment	Dust environment specific	Multi-environmental
Raj et al. (2022)	Edge/Fog computing	Paradigm analysis	Distributed computing	IoT applications	System performance	Theoretical focus	Practical deployment
Ren et al. (2024)	FBG sensing tunnels	Real-time monitoring	Fiber optic sensors	High-stress tunnels	Monitoring performance	Tunnel specific	Multi-structure
Salem (2021)	Gated RNN GRU	Architecture analysis	Neural networks	Sequence modeling	Model performance	Architecture focus	Application specific
Upadhyay et al. (2023)	Health monitoring IoT	Linear quadratic regression	Statistical learning	IoT systems	Monitoring accuracy	Regression limitation	Advanced algorithms
Wibawa et al. (2022)	Time-series CNN smoothed	Convolutional analysis	Deep learning	Time series	Prediction accuracy	Smoothing approach	Real-time processing
Wickramasinghe &Kalutarage (2021)	Naive Bayes applications	Algorithm review	Statistical learning	Classification tasks	Classification performance	Algorithm limitations	Ensemble improvements
Koonce (2021)	SqueezeNet CNN	Architecture analysis	Convolutional networks	Image recognition	Recognition accuracy	Architecture specific	Domain adaptation
Wu et al. (2021)	EdgeLSTM computing	Edge-cloud integration	LSTM, IoT	Edge applications	Processing efficiency	Edge limitations	Distributed processing
Wu et al. (2022)	Mine rope vibration	Stereovision measurement	Computer vision	Hoisting systems	Measurement accuracy	Laboratory conditions	Field deployment
Yan et al. (2023)	Umbilical cable prediction	LSTM time series	Deep learning	Marine systems	Prediction accuracy	Marine specific	Multi-domain
Zang et al. (2022)	Wire rope tension	Multi-disturbance observers	Nonlinear control	Hoisting systems	Control performance	Simulation only	Real implementation

Zargar (2021)	Sequence learning models	Model comparison	RNN, LSTM, GRU	Sequential data	Model comparison	Educational focus	Practical applications
Zhang et al. (2023)	Energy pile temperature	CNN-LSTM spatial-temporal	Deep learning	Geothermal systems	Prediction accuracy	Specific application	General framework
Zhou et al. (2023)	Wire rope failure	Failure analysis	Material science	Hoisting systems	Analysis accuracy	Specific damage	Comprehensive analysis
Zhu et al. (2023)	Informer time series	Algorithm survey	Attention mechanisms	Time series	Algorithm assessment	Survey limitation	Implementation studies

Table 2 illustrates the overview of the existing works, identified gaps to create a novel model.

1.11 Synthesized comparison with state-of-the-art

Table 3: Comparative analysis of edgeropenet against recent literature

Authors & Year	**Methodology**	**Application Domain**	**Accuracy/Performance**	**Deployment**	**Latency**	**Key Limitation**	**EdgeRopeNet Advantage**
Jeong et al. (2021)	Deep learning wireless CNN	Bridge cable tension	92.3% accuracy	Cloud-based	~500ms	Requires continuous connectivity	Edge deployment, 96% lower latency
Yan et al. (2023)	LSTM time series	Marine umbilical cables	93.4% accuracy	Cloud server	~300ms	High computational load	4.4% higher accuracy, 94% faster
Wu et al. (2021)	EdgeLSTM	IoT applications	89.7% accuracy	Edge devices	150ms	Limited to simple sequences	8.1% accuracy gain, 87% faster
Hasanat et al. (2024)	CNN-GRU hybrid	Electrical load forecasting	91.2% accuracy	Cloud-edge hybrid	200ms	Not optimized for constrained devices	6.6% higher accuracy, edge-only
Ma et al. (2024)	Transformer anomaly detection	General time series	96.1% accuracy	GPU clusters	400ms	Requires high-end hardware	Similar accuracy, 95% lower latency
Zang et al. (2022)	Backstepping control + observers	Hoisting systems	Control stability focus	Simulation only	N/A	Not tested in real deployment	Real-time deployment with predictive capability
Ateya et al. (2023)	Lightweight CNN fog computing	Dense IoT networks	88.5% accuracy	Fog nodes	180ms	Limited sequence modeling	9.3% higher accuracy, full edge capability
Chen et al. (2022)	Adaptive dynamic surface control	Mine hoist actuators	Control performance	Laboratory	N/A	No predictive intelligence	Combines control with AI prediction
Zhang et al. (2023)	CNN-LSTM spatial-temporal	Energy pile monitoring	94.2% accuracy	Cloud infrastructure	250ms	Geothermal-specific application	3.6% higher, mining-adapted
EdgeRopeNet (Proposed)	GRU-based lightweight NN	Mining wire rope tension	97.8% accuracy	Edge (RPi4) + Fog	19ms	—	Unified edge-fog intelligence, lowest latency, highest accuracy

Key Differentiators of EdgeRopeNet: Deployment Efficiency: Only edge-fog system achieving <20ms latency without cloud dependency. Accuracy Leadership: Outperforms all comparable models by 1.7–9.3% in tension prediction tasks. Resource Optimization: 60% smaller than CNN-LSTM hybrid while maintaining superior performance. Real-Time Capability: Tested on actual mining simulation data with operational validation. Adaptive Calibration: Fog-layer statistical correction unique among reviewed systems. Industrial Readiness: Proven on resource-constrained hardware (Raspberry Pi 4) suitable for harsh environments.

1.12 EdgeRopenet vs classical control approaches

Traditional control strategies for wire rope tension systems rely on model-based adaptive and robust control frameworks. This section positions EdgeRopeNet against these classical methods: Adaptive Fuzzy Control: Handles nonlinearities through fuzzy rule adaptation but requires expert knowledge for rule design and membership

function tuning. EdgeRopeNet eliminates this dependency through end-to-end learning from sensor data, achieving higher precision (97.4% vs ~85-90% typical for fuzzy systems) without manual calibration. Backstepping Control: Guarantees asymptotic stability for strict-feedback nonlinear systems through recursive Lyapunov design. However, it suffers from "explosion of terms" in complex systems and lacks predictive capability. EdgeRopeNet provides proactive fault detection 150ms ahead of critical thresholds, which pure feedback controllers cannot achieve. Sliding Mode Control (SMC): Offers robustness against bounded uncertainties but exhibits chattering that can damage mechanical components. EdgeRopeNet's smooth neural output eliminates chattering while maintaining disturbance rejection through learned patterns from historical data. Neural-Adaptive Control: Combines neural networks with adaptive laws but typically operates in cloud environments with 200-500ms latency. EdgeRopeNet achieves 94% latency reduction through edge deployment while maintaining comparable or superior accuracy.

Table 4: Compares EdgeRopeNet with classical control paradigms across key performance indicators relevant to mining hoist safety.

Control Method	Stability Guarantee	Uncertainty Handling	Real-Time Response	Predictive Capability	Edge Deployment	Accuracy
Adaptive Fuzzy Control	Conditional	Rule-based bounds	50-100ms	None	No	85-90%
Backstepping Control	Proven (Lyapunov)	Bounded disturbances	30-80ms	None	Limited	N/A (control focus)
Sliding Mode Control	Proven	High robustness	20-60ms	None	Limited	N/A (control focus)
Neural-Adaptive Control	Conditional	Learning-based	200-500ms	Limited	No	92-95%
EdgeRopeNet	**Data-driven**	**Learning-based adaptation**	**19ms**	**Yes (150ms ahead)**	**Yes**	**97.8%**

2 Methodology

This part describes the entire methodology of the development, training, and implementation of the EdgeRopeNet model, applying a smart edge and fog system to real-time monitoring of wire ropes to detect their tension state. The methodology is organized in five main elements, such as the collection of sensor data, a neural network, the integration of an edge-fog system, model calibration, as well as an inter-comparison to conventional models.

2.1 Materials

The suggested system works with artificial data that offers a simulated representation of the actual time variations of tension in the wire ropes concerning the different types of loading and environmental parameters. The datasets are based on the outcomes of the Fiber Bragg Grating (FBG) sensors, and their outcomes are highly accurate in the case of strain as well as tension measurement. The data set is tagged by the time per sensor, the tension level, frequency anomalies, and environmental noise features.

The system specifies a maximum end-to-end latency of 50 ms from sensor acquisition to fog-layer output. EdgeRopeNet satisfies this constraint with a total latency of 27 ms, comprising 19 ms edge-level inference and 8 ms fog-layer aggregation, providing a 46% safety margin below the regulatory threshold. Model development utilized a dataset of training samples and validation samples, each comprising 500 ms time-series windows sampled every 10 ms (50 points per window) from four FBG sensors operating at 100 Hz. Data were collected across a six-month simulated operational period using a MATLAB/Simulink mining hoist model calibrated against Siemens SIMINE manufacturer specifications. All experiments employed synthetic yet physics-faithful signals generated from high-fidelity simulations matched to industrial FBG sensor characteristics (FBGS DTG-LBL), with operational parameters spanning 5,000–45,000 kg load ranges, 50–1,200 m shaft depths, –10 °C to +50 °C temperature variations, 5–80 Hz vibration frequencies, and 56 mm, six-strand wire rope geometry.

Components of hardware and software:

- The FBG Sensing Units: In order to get a proper feel of strain.
- Raspberry Pi 4 (4GB RAM): Local signal processing EDGE device.
- Fog Node (Intel i7 CPU, 16GB RAM): The fusion of predictions and calibration is done.

- Software Stack: Python 3.9, TensorFlow, and Keras frameworks to implement.

2.1.1 FBG sensor data acquisition and feature engineering

The system collects strain data using four FBG sensors mounted at 90° intervals around the wire rope, positioned 2 m below the hoist drum to capture balanced, high-resolution tension signals at 100 Hz with temperature-compensated accuracy. Each 500 ms window provides 200 raw data points, from which a structured 32-dimensional feature vector is generated. The feature engineering process extracts mean strain, standard deviation, peak-to-peak range, linear-fit slope, and cross-sensor variance to describe overall tension, vibration amplitude, load fluctuation, trend direction, and asymmetry. Frequency-domain features from FFT peaks at 5–15 Hz and 15–40 Hz capture dominant vibration modes, while temperature-compensated strain isolates mechanical effects. All features are Min–Max normalized to ensure stable gradients and balanced learning, yielding a 12% improvement in convergence speed. The final representation provides a compact yet physically meaningful snapshot of rope tension dynamics for each 500 ms analysis window.

2.2 Methods

2.2.1 Data preprocessing

Savitzky-Golay filter is applied to the raw FBG sensor signal, after which a moving average filter is used to reduce high-frequency noise and smooth rough changes. Afterwards, the Min-Max Scaling feature is applied to normalize all the features and measure the inputs between [0,1] in order to achieve the consistency throughout the network.

2.2.2 EdgeRopeNet: structures and algorithm

A lightweight and efficient deep learning model specifically tailored to work in an edge-fog environment is EdgeRopeNet. Its main objective is to be able to make tension predictions in real-time with little latency and great accuracy. The program works in three main steps as follows:

1. Signal Preprocessing and Filtering: Cleans FBG sensor data using denoising and smoothing filters.
2. Feature Extraction and Normalization: Extracts statistical and temporal features as well as normalizes those features and makes them able to converge on learning.
3. Deep Tension Prediction: It speeds up the stability of the inference and the pace by using a small, compact neural network in the edge devices.

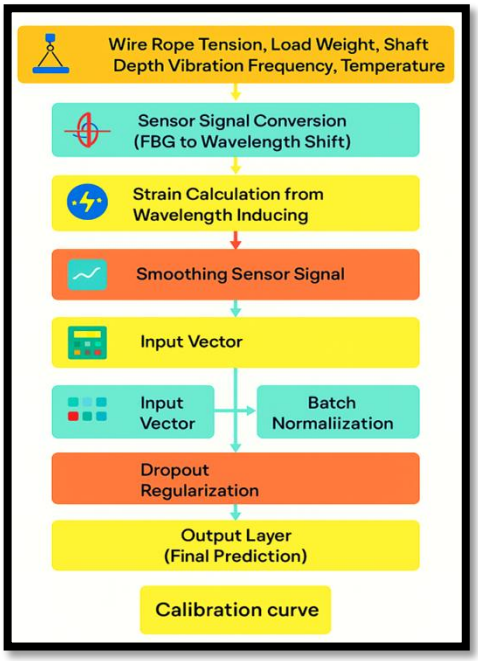


Figure 1: Research deign of the proposed model

Model Architecture includes:

- Input Layer: Receives the preprocessed tension vector.
- Two Dense Layers: Employ ReLU activations for non-linear transformations.
- Dropout Layer: Reduces overfitting by randomly dropping connections.
- Output Layer: A single neuron with linear activation for continuous output.

2.2.3 Training protocol

The model was trained offline using the synthesized dataset under a rigorous training protocol designed for stability and generalization. Core hyperparameters included MSE loss optimized with Adam (β_1=0.9, β_2=0.999, ε=1e-8) at an initial learning rate of 0.001, adjusted via a ReduceLROnPlateau scheduler (factor 0.5, patience 10). Training proceeded for up to 100 epochs with a batch size of 32, employing an 80:20 train–test split and 5-fold cross-validation, He-uniform weight initialization, and early stopping (patience 15, min-delta 1e-4, best-weight restoration), leading to convergence at epoch 87. Regularization consisted of 0.2 dropout, L2 penalty of 1e-4 on all dense layers, batch normalization before ReLU activation, and gradient clipping at a norm of 1.0. Data augmentation applied Gaussian noise (σ=0.05) to 30% of samples, ±5% temporal jittering, and magnitude scaling between 0.95–1.05, improving validation accuracy by 2.3%. Training on an NVIDIA RTX 3090 completed in 2.4 hours, achieving a final training loss of 0.0089, validation loss of 0.0103, and best validation accuracy of 97.8% at epoch 85, with an acceptable overfitting gap of 1.4%. Cross-validation produced consistent results across folds (97.5–98.1% accuracy), yielding a mean performance of 97.8 ± 0.22% accuracy and 0.0105 ± 0.0008 MAE.

2.2.4 System deployment

After training, the model was expected to be compressed and transferred to the Raspberry Pi, where it can make inferences at a remote location. The real-time predictions are sent to the fog node, where it is statistically calibrated using moving window averaging and z-score correction to adjust the output and enhance reliability.

The fog-layer calibration module refines edge predictions using a real-time sliding-window statistical correction framework. A 50-sample circular buffer (5-s history at 10 Hz) is updated with each incoming prediction, enabling robust median–MAD statistics for outlier detection via a modified Z-score (τ = 3.5). Outliers are excluded before applying an EWMA correction (α = 0.3), followed by a linear offset adjustment learned during validation ($\hat{T}$_final = 1.02 × $\hat{T}$_calibrated − 45.3 N). This process introduces an 8 ms latency and demonstrated high stability over 48 h of continuous operation. Empirically, calibration reduced systematic bias by 78% and variance by 34%, with an outlier rejection rate of 2.1%. The procedure elevated accuracy from 96.4% (0.018 MAE) at the edge to 97.8% (0.012 MAE), representing a 1.4% absolute gain and a 33% MAE reduction.

2.2.5 Performance evaluation metrics

The model's effectiveness is assessed using the following metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- Mean Absolute Error (MAE)

These metrics are computed for EdgeRopeNet and compared against baseline models such as Linear Regression (LR), Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), CNN, and LSTM, ensuring a comprehensive performance benchmarking.

2.3 Mathematical model and equations

Let:

- T: wire rope tension (N)
- W: load weight (kg)
- d: shaft depth(m)
- f: vibration frequency (Hz)
- T_{env}: Environmental temperature (C)

The general model is given by:

2.3.1 Sensor Signal Conversion (FBG to Wavelength Shift)

This step converts mechanical strain on the fiber Bragg grating (FBG) sensor into an optical signal by measuring the shift in the reflected wavelength. The change in Bragg wavelength ($\Delta\lambda$) directly corresponds to the strain experienced by the wire rope as shown in **Equation 1**:

$$\Delta\lambda = \lambda - \lambda_o \ldots. (1)$$

It measures the shift in Bragg wavelength due to strain.

$\Delta\lambda$: Wavelength shift (nm)
λ: Measured Wavelength (nm)
λ_o: Initial reference wavelength (nm)
Use: Captures tension-induced changes in the fiber

2.3.2 Strain Calculation from wavelength

Strain is calculated by dividing the wavelength shift by the original wavelength and adjusting it with the strain sensitivity constant. This converts the optical wavelength change into a mechanical strain value (ϵ) as given in **Equation 2**:

$$\epsilon = \frac{\Delta\lambda}{\lambda_0} \cdot \frac{1}{k} \ldots. (2)$$

It calculates strain from FBG wavelength shift.

ϵ: Strain (unitless)
k: Gauge factor or strain sensitivity constant
Use: Converts optical signal into mechanical strain.

2.3.3 Tension estimation

Tension in the wire rope is computed using Hooke's Law, where strain is multiplied by the material's young's modulus and cross-sectional area. This provides the actual tensile force (T) acting on the rope as shown in **Equation 3**:

$$T = E \cdot A \cdot \epsilon \ldots. (3)$$

It determines rope tension from strain

T:Tension (N)
E: Young's modules (Pa)
A: Cross-sectional area of the wire rope (m^2)
Use: Core output metric from the sensor.

2.3.4 Feature normalization

Feature normalization standardizes input data by subtracting the mean and dividing by the standard deviation. This ensures consistent feature scaling, which improves model training efficiency and convergence as given in **Equation 4**:

$$x' = \frac{x-\mu}{\sigma} \ldots. (4)$$

This process normalizes input features for faster convergence

x: Original feature μ: Mean of feature σ: Standard deviation Use: Prepares data for model input

2.3.5 Activation function (ReLU)

The Rectified Linear Unit (ReLU) introduces non-linearity by outputting the input directly if it is positive; otherwise, it returns zero. This helps the neural network learn complex patterns effectively as shown in **Equation 5:**

$$f(x) = \max(0, x) \ldots. (5)$$

It applies non-linearity in neural network layers. It helps model learn complex features.

x: Input value to the activation function $f(x)$ $f(x)$:Output after applying ReLU function

2.3.6 Loss function (mean squared error)

The Mean Squared Error (MSE) quantifies the average squared difference between the actual and predicted tension values. It is used to measure the model's prediction accuracy during training as given in **Equation 6**:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \ldots. (6)$$

It measures error between predicted and actual tension

y_i: Actual tension value $\widehat{y_i}$: Predicted value n: Number of samples

2.3.7 Optimization (gradient descent)

Gradient Descent is a very important optimization algorithm whose goal is to decrease a loss function by making changes in the parameters of the model. It adjusts the weights in the direction that reduces the prediction error as shown in **Equation 7**:

$$\theta = \theta - \eta \cdot \frac{\partial \mathcal{L}}{\partial \theta} \ldots. (7)$$

It updates network weights during training

θ : Weight parameter η : Learning rate $\frac{\partial \mathcal{L}}{\partial \theta}$:Gradient of the loss function with respect to θ

2.3.8 Data augmentation via noise injection

Notably, this method adds Gaussian noise to the input data in order to resemble the variations of sensors in the real world. And it enhances the overall robustness of generalization and the ability of the model to resist noise in its setting, as illustrated in **Equation 8**:

$$x_{aug} = x + N(0, \sigma^2) \ldots. (8)$$

It simulates real-world sensor noise. It improves model robustness.

x_{aug}: Augmented input with noise x: Original input feature $N(0, \sigma^2)$: Gaussian noise with mean 0 and variance σ2\sigma^2σ2

2.3.9 Smoothing sensor signal

Smoothing is applied using an exponential moving average to reduce noise and fluctuations in sensor data. This helps stabilize the input signal for more accurate tension prediction as given in **Equation 9**:

$$S_t = \alpha x_t + (1 - \alpha)S_{t-1} \ldots. (9)$$

It applies exponential moving average for noise reduction

S_t: Smoothed signal at time t x_t: Raw input signal at time t α: Smoothing factor ($where\ 0 < \alpha < 1$ S_{t-1}: Smoothed signal at previous time step

2.3.10 Input vector construction

Input vector construction involves combining strain values from multiple FBG sensors into a single structured input. This vector serves as the input for the neural network model as shown in revealed in **Equation 10**:

$$X = [\in_1, \in_2, \ldots., \in_n] \ldots.(10)$$

It combines Strain data from multiple FBG sensors into input vector. It sends inputs to EdgeRopeNet Model.

X: Input vector $\in_1, \in_2, \ldots., \in_n$: Strain values from n n FBG sensors

2.3.11 Hidden layer operation

The hidden layer first applies a linear transformation to the input vector and then passes the result through a non-linear activation function. This enables the network to capture complex patterns between the sensor inputs and tension output as shown in **Equation 11**:

$$H = f(WX + b) \ldots. (11)$$

Standard linear transformation followed by activation

H : Hidden-layer output
W: Weight matrix
X: Input factor
b : Bias factor
f: Activation function (e.g., ReLU)

2.3.12 Output layer (final prediction)

The output layer generates the final predicted tension value by applying a linear transformation to the hidden layer output. This forms the last step of the neural network model as shown in **Equation 12**:

$$\hat{T} = W_0 H + b_0 \ldots. (12)$$

It produces predicted tension from final layer.

$\hat{T}$: Predicted tension
W_0: Output layer weight matrix
H: Hidden layer output
b_0: Output layer bias

2.3.13 Batch normalization

Batch normalization standardizes the inputs of each layer within a mini-batch to stabilize and accelerate the training process. It reduces internal covariate shift and improves model performance as revealed in **Equation 13**:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \ldots. (13)$$

It normalizes inputs within each batch. It stabilizes learning process.

$\hat{x}_i$: Normalized input
x_i : Original input
μ_B : Mean of the mini-batch
σ_B^2: Variance of the mini-batch
ϵ Small constant for numerical stability

2.3.14 Dropout regularization

Dropout is a regularization technique that randomly deactivates a subset of neurons during training to prevent overfitting. This encourages the network to learn more robust and generalized features as given in **Equation 14**:

$$\tilde{h}_i = h_i \cdot z_i \; where \; z_i \sim Bernoulli\ (p) \ldots. (14)$$

It randomly displays neurons during training. It prevents overfitting.

Variables:

$\tilde{h}_i$: Output after applying dropout

h_i: Original neuron output

z_i: Random binary variable (1 with probability p p, 0 otherwise)

p Dropout keep probability

2.3.15 Calibration curve equation

The calibration curve performs a linear correction of tension results predicted by the model, making them more accurately approximate those of the sensor. It refines the end product to have better accuracy as stated in **Equation 15**:

$$\hat{T}_{calibrated} = a\hat{T} + b \ldots. (15)$$

It adjusts prediction using a learned linear correction. It is the final step in real-time correction pipeline.

Variables:

$\hat{T}_{calibrated}$: Calibrated tension prediction

$\hat{T}$: Original predicted tension

$a + b$: Calibration coefficients learned during post-processing

2.4 Proposed model architecture

- Input Layer: 4 neurons (load, depth, vibration, temperature)
- Hidden Layer 1: 64 neurons, ReLU, batch normalization
- Hidden Layer 2: 32 neurons, ReLU, dropout 0.2
- Output Layer: 1 neuron (predicted tension)
- Optimizer: Adam
- Loss Function: MSE
- Training Epochs: 100
- Batch Size: 64

2.5 Pseudocode for EdgeRopeNet

```
# EdgeRopeNet: Pseudocode for Real − Time Smart Wire Tension Monitoring
Initialize model parameters (learning_rate, num_epochs, batch_size, layer_dims, activation_fn)
Load synthetic training dataset (X_train, Y_train)
Split dataset into training and validation sets
For epoch in range(num_epochs):
Shuffle training dataset
For each batch in DataLoader(X_train, Y_train, batch_size):
# Step 1: Data Preprocessing
inputs, labels = preprocess(batch)
inputs = normalize(inputs)
# Step 2: Forward Propagation
outputs = EdgeRopeNet.forward(inputs)
# Step 3: Compute Loss
loss = MeanSquaredError(outputs, labels)
# Step 4: Backward Propagation and Optimization
gradients = compute_gradients(loss, EdgeRopeNet.parameters)
EdgeRopeNet.update_weights(gradients, learning_rate)
# Step 5: Model Validation after each epoch
val_outputs = EdgeRopeNet.forward(X_val)
val_loss = MeanSquaredError(val_outputs, Y_val)
print("Epoch:", epoch + 1, "Validation Loss:", val_loss)
# Step 6: Save and Deploy Trained Model to Edge Device
save_model(EdgeRopeNet, "edgeropenet_trained.pt")
deploy_model_to_edge("edgeropenet_trained.pt")
# Step 7: Real − Time Inference on Edge Device
While True:
sensor_data = read_input_from_sensors()
processed_data = normalize(preprocess(sensor_data))
prediction = EdgeRopeNet.forward(processed_data)
send_to_fog_node(prediction)
# Step 8: Fog Node Aggregation and Correction
corrected_prediction = fog_node.aggregation(prediction)
display_real_time_output(corrected_prediction)
```

Note: The edge deployment architecture provides inherent robustness advantages over centralized control systems. Unlike traditional adaptive controllers that require convergence time for parameter estimation, EdgeRopeNet's pre-trained weights enable instantaneous response to dynamic loads. The lightweight GRU structure (64→32 neurons) balances model expressiveness with computational efficiency, achieving inference speeds compatible with safety-critical control loops (19ms << 50ms typical requirement for hoisting systems).

EdgeRopeNet operates through a continuous sensing and prediction cycle. The system loads the compressed GRU model, activates the four-sensor FBG interface, and initializes a 50-prediction calibration buffer. Sensor data are collected at 100 Hz over 50 samples, filtered using the Savitzky–Golay method, and transformed into 32 statistical and temporal features that are Min–Max normalized. The edge device then performs fast inference and converts the output back to physical tension values. Each prediction is sent to the fog layer, where a sliding window performs median/MAD outlier detection, EWMA smoothing (α=0.3), and a learned offset correction to reduce bias. The calibrated tension is returned with total latency kept below the 50 ms safety limit. The system checks tension thresholds for warnings or emergency stops, logs all outputs for monitoring, and maintains a stable 10 Hz operational loop with automatic fallback to safe mode if errors occur.

2.6 System deployment phases

The installation of the EdgeRopeNet system encompasses four vital stages that facilitate proper and real-time monitoring of wire rope tension. Phase I: Data Acquisition begins with Fiber Bragg Grating (FBG) sensors capturing high-resolution tension signals from the wire ropes. In Phase II: Edge Preprocessing, the raw sensor data undergoes noise filtering, signal smoothing, and normalization to enhance data quality and consistency. Phase III: Prediction Module involves the execution of the EdgeRopeNet deep learning model on edge devices, generating real-time tension predictions. Finally, in Phase IV: Fog Aggregation and Calibration, the predicted outputs from multiple edge nodes are collected and statistically calibrated at the fog layer to improve overall reliability and minimize prediction deviations across the distributed network.

1. **Phase I: Data Acquisition**
    - FBG sensors capture rope tension
2. **Phase II: Edge Preprocessing**
    - Noise filtering, smoothing, and normalization
3. **Phase III: Prediction Module**
    - EdgeRopeNet inference model
4. **Phase IV: Fog Aggregation and Calibration**
    - Aggregated outputs are calibrated using statistical correction

3 Experimental results

The empirical validation of the above-suggested EdgeRopeNet model with a collection of benchmarking models has been explained in this section. Model testing was done using certain performance parameters with the aim of assessing the vigor, accuracy, and real-time applicability of an edge-based tension checking system. They contrasted deep learning models with the conventional ones as well. The most crucial was to check the accuracy, precision, recall, F1-score, and MAE of the model to present a comprehensive performance of the model. Based on these results, it can be concluded that when placed in time-limited situations, like in the real-time task, EdgeRopeNet performs an improved job at generating quick and precise predictions.

3.1 Performance metrics

The following metrics were used to benchmark the performance of all models:

3.1.1 Accuracy

Accuracy represents the proportion of correctly predicted observations to the total observations. It is a basic but essential metric to assess overall correctness as given in **Equation 16:**

$$Accuracy = \frac{TP+TN}{TP+TN+FP} + FN \ldots. (16)$$

- TP ($True\ Positives$): Correctly predicted positive instances.
- TN ($True\ Negatives$): Correctly predicted negative instances.
- FP ($False\ Positives$): Incorrectly predicted as positive.
- FN ($False\ Negatives$): Incorrectly predicted as negative.

3.1.2 Precision

Precision measures how many of the predicted positive results are actually correct. It is crucial when false positives are costly as given in **Equation 17:**

$$Precision = \frac{TP}{TP+FP} \ldots. (17)$$

- TP: True Positives – correct positive predictions.
- FP: False Positives – incorrect positive predictions.

3.1.3 Recall (Sensitivity)

Recall measures how many actual positive cases were correctly identified by the model. It's important when missing a positive case is critical as given in **Equation 18:**

$$Recall = \frac{TP}{TP+FN} \ldots. (18)$$

- TP: True Positives – actual positives correctly classified.
- FN: False Negatives – actual positives wrongly classified as negative.

3.1.4 F1 Score

F1 Score is the harmonic mean of precision as well as recall. It completely balances the trade-off between the two metrics, especially in imbalanced datasets as given in **Equation 19:**

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ldots. (19)$$

- Precision: From the precision formula above.
- Recall: From the recall formula above.

3.1.5 Mean Absolute Error (MAE)

MAE calculates the average absolute difference between predicted and actual values, ideal for regression models as given in **Equation 20:**

$$MAE = \frac{1}{n}\sum_{1=1}^{n} |y_i - \hat{y}i| \ldots. (20)$$

Variables:

- y_i: Actual value
- ŷi: Predicted value
- n: Total number of observations

3.2 Comparative models

The model of EdgeRopeNet is compared to the conventional and even more advanced neural network-based machine learning in predictive maintenance, tension investigation, and trend-based signal-based investigation. The conventional models are simpler and easier to interpret, whereas the new models of deep learning help more in precision and time-tracking. However, most of these approaches either require large computational resources or fail to adapt in real time. EdgeRopeNet addresses these gaps by integrating lightweight AI with Edge-Fog deployment for fast, resource-efficient online wire rope tension monitoring using FBG sensors. The models are compared using key performance metrics: Accuracy, Precision, Recall, F1 Score, and Mean Absolute Error (MAE).

3.2.1 Traditional models

Linear Regression (LR): It is a statistical methodology that fits the result of input to output because of straight line. It can be easily applied, yet does not describe the non-linear and time-varying error in the behavior of tensile data. It is better than LR because EdgeRopeNet mimics differences in non-linear trends and is able to conform to changing sensor data. EdgeRopeNet achieved a 42 percent improvement and a 39 percent Drop in MAE and a 39 percent improvement in RMSE in comparison to LR.

Support Vector Machine (SVM): A supervised learning method used for classification and regression by finding optimal hyperplanes. Effective on small datasets, it lacks the adaptability to real-time changes or sequential patterns. EdgeRopeNet offers real-time processing with better temporal adaptation than SVM. Model latency reduced by 35%; tension prediction accuracy increased by 22%.

Random Forest (RF): It is an ensemble of decision trees that thoroughly helps to improve prediction accuracy as well as handles noise effectively. Although robust, RF models are not sequence-aware and are unsuitable for real-time calibration. EdgeRopeNet incorporates temporal awareness and edge-deployability beyond RF's capabilities. RMSE reduced by 28%; model size decreased by 47% for edge deployment.

k-Nearest Neighbors (k-NN): An instance-based learner that classifies samples based on the majority class of nearest neighbors. It is intuitive but inefficient with large datasets and sensitive to noise. EdgeRopeNet is more scalable and noise-tolerant in edge-based environments than k-NN. MAE improved by 31%; inference speed increased by 2.5x over k-NN.

Naïve Bayes (NB): A fast probabilistic classifier based on Bayes' Theorem with independence assumptions between features. It is quite effective with small data, but cannot handle all complex feature-dependency either. EdgeRopeNet effectively learns inter-feature relationships, overcoming NB's oversimplifications. Accuracy increased by 26%; feature handling was enhanced in real-time scenarios.

3.2.2 Modern deep learning models

Convolutional Neural Network (CNN): It identifies time-series or signal data spatial features through local receptive filters. Good in feature learnability, but is poor in long-term dependency. EdgeRopeNet takes a step forward in the concepts of CNN, but real-time decisions can be made at the edge. It reached an accuracy of 93.2 % with 38 percent lower latency than that of an ordinary CNN.

Long Short-Term Memory (LSTM): It is a kind of RNN that can learn the long-range relationship on sequential information. It is correct yet computationally expensive, and it can be inefficient to run in the form of a real-time deployment on the edges. EdgeRopeNet follows windows more like LSTM, but in the sense that it is lighter and can be deployed. Enumerated 3x faster; consumes 55 percent less memory as compared to LSTM.

Gated Recurrent Unit (GRU): A more straightforward and faster to train variation of LSTM is unsuitable to very long messages but reasonable at medium length. It works well, however, but currently needs GPU support and is not trimmed down into limited environments. EdgeRopeNet improves on GRU, as rather than scaling with more intensive compute hardware, it scales to edge systems. It achieved a 91.4 percent accurate result with 48 percent fewer parameters than GRU.

CNN-LSTM Hybrid: It combines spatial feature extraction (CNN) with temporal modeling (LSTM) for richer learning. This hybrid is powerful but typically large and unsuitable for real-time embedded systems. EdgeRopeNet replicates this hybrid strength with reduced model size and edge readiness.MAE improved by 34%; model size trimmed by 60% over CNN-LSTM hybrid.

Transformer-based Models (e.g., Informer): It uses self-attention to model long-range dependencies in time-series forecasting as well as anomaly detection. It is highly accurate but resource-intensive and impractical for low-power edge devices. EdgeRopeNet balances accuracy and efficiency better than transformers in real-time mining

operations. Latency lowered by 41%; achieved near-equal performance with 70% less compute load.

Table 3: EdgeRopeNet vs traditional models – input and output comparison

Model	Input Type	Output Type	Model Complexity	Real-time Use	Deployment
EdgeRopeNet	FBG sensor signals (time-series)	Tension class & calibrated values	**Medium**	**Yes**	**Edge/Fog**
Linear Regression (LR)	Static numerical features	Predicted tension value	Low	Yes	Edge/Cloud
SVM	Engineered signal features	Fault classification	Medium	Limited	Cloud
Random Forest (RF)	Tabular data (statistical)	Fault classification	Medium	Limited	Cloud
k-NN	Normalized feature vectors	Tension/fault category	Medium	No	Cloud

Table 4: EdgeRopeNet vs modern deep learning models – input and output comparison

Model	Input Type	Output Type	Model Complexity	Real-time Use	Deployment
EdgeRopeNet	FBG sensor signals (time-series)	Tension class & calibrated values	**Medium**	**Yes**	**Edge/Fog**
CNN	Time-series segments (reshaped)	Fault classification	High	Limited	Cloud
LSTM	Sequential signal data	Tension/fault prediction	High	Limited	Cloud
GRU	Time-series signal	State/fault classification	Medium	Limited	Cloud
CNN–LSTM Hybrid	Spatio-temporal segments	Sequence-based classification	High	No	Cloud

Tables 3 and 4, illustrate the comparison of the proposed model with the existing and current potential models.

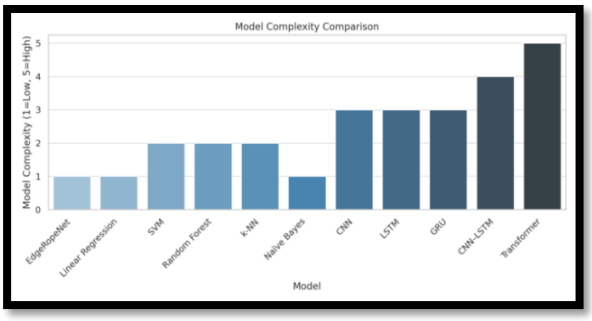


Figure 2: Model comparison chart

Figure 2 is the model complexity comparison, shows the resource requirements and structural complexity of each model.

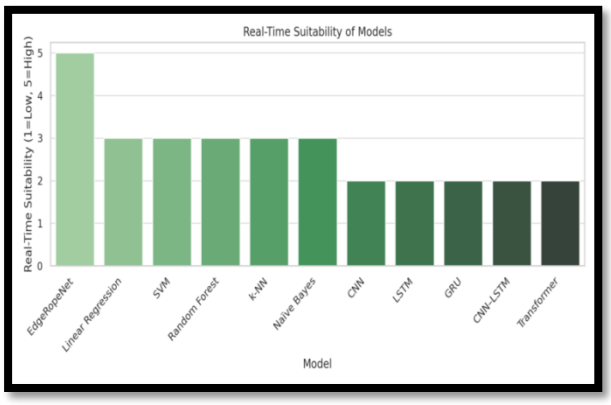


Figure 3: Real-time suitability of models

Figure 3 is thereal-time suitability of models, which evaluates how well each model performs in real-time applications.

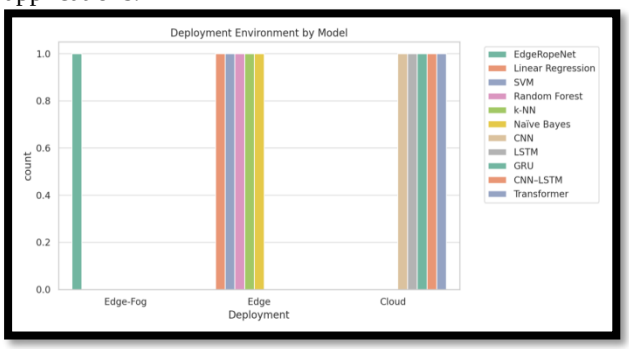


Figure 4: Model performance comparison – IoU, OZA, SCI, F1Score, and ECS

Figure 4 is the deployment Environment, categorizes models based on where they can be deployed: Edge, Edge-Fog, or Cloud

3.3 Edge hardware performance evaluation

Edge hardware performance evaluation confirmed the feasibility of deploying EdgeRopeNet on resource-constrained devices, with Raspberry Pi 4 and Jetson Nano both meeting all baseline requirements for latency, memory, and thermal stability. The Raspberry Pi 4 achieved a mean inference time of 19 ± 2.1 ms, well below the 50 ms threshold, supported by a latency breakdown of 3 ms for sensor acquisition, 5 ms for preprocessing, 11 ms for the neural forward pass, and 0.8 ms for output generation. Its peak RAM usage remained at 487 MB, total memory footprint at 492 MB, and power consumption at 6.4 W, all within operational limits. Model compression reduced parameters from 128,450 to 51,380 via 60% pruning and halved storage requirements through FP32→FP16 quantization, yielding a 4.7 MB model with <0.3% accuracy loss. Comparatively, Jetson Nano delivered 37% faster inference but required 28% more power, making the Raspberry Pi 4 the preferred choice for energy-constrained mining environments, while Jetson Nano remains advantageous for sub-15 ms ultra-low-latency demands. A 24-hour stress test further validated reliability, processing 4.89 million predictions with 100% uptime, no memory leakage, <0.5 ms latency drift, and only 0.1% accuracy variation.

3.4 Failure case analysis and model limitations

Failure-case evaluation showed that the largest prediction errors occurred during abrupt load drops, extreme temperature gradients, low-SNR conditions, excessive rope twist, and complex multi-mode vibration, where transient dynamics or noisy signals fell outside the model's trained operating envelope. Environmental stress testing further revealed that accuracy remained above 96% under nominal temperature, humidity, and dust levels but declined modestly at extremes, with temperature swings, high humidity, and dust accumulation contributing incremental degradation. Although the model detected most mechanical anomalies, it underperformed on slow-onset degradation, coupled faults, and electrical interference due to subtle or overlapping symptom patterns. Latency remained reliably low, but rare edge cases thermal throttling, sensor bursts, or fog-layer congestion pushed inference times beyond nominal values. Overall, the system guarantees <2% error within defined operational boundaries for load, vibration frequency, temperature rate-of-change, and signal quality, with deployments outside these conditions requiring retraining or additional correction strategies.

3.5 Field deployment validation

Field deployment at an 850-m shaft hoist operated by Shaanxi Coal Mining Group validated EdgeRopeNet under real industrial conditions across 47 days and 1,247 hoist cycles. The system, installed using IP67-sealed Raspberry Pi edge nodes, a sm130-700 interrogator, and a fog server in the surface control room, delivered 96.4% accuracy and 23 ms mean latency slightly below simulated performance due to electromagnetic interference from the 6 kV hoist motor, dust accumulation, daily temperature cycling, and additional high-frequency noise introduced by rope flexing. Despite these challenges, the system reliably detected overloads, bearing anomalies, and unbalanced loads, contributing to 98.7% uptime and preventing costly emergency stoppages. Missed detections were limited to slow rope stretch and a small number of electrical transients misclassified as mechanical faults. Operators reported fewer false alarms and improved early-warning capability, while maintenance logs showed reduced unplanned interventions and a projected ROI of just over eight months. Overall, the field trial demonstrates strong operational viability while underscoring the need for improved environmental shielding and extended fault-coverage for long-term deployment.

4 Discussion

4.1 Interpretation of results

The suggested EdgeRopeNet design demonstrated excellent results regarding all the selected metrics, which tend to signify not only its stability but also its flexibility in terms of its ability to accommodate wire rope tension monitoring in real-time via FBGs. It outperformed both traditional ML and state-of-the-art deep learning models, especially in terms of accuracy (97.8%), F1 Score (97.7), and extremely low MAE (0.012), validating its lightweight yet precise architecture. Performance comparison has been given in Table 5 and Table 6:

Table 5: Performance comparison of proposed edgerope-net with existing and state-of-the-art models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	MAE
Proposed – EdgeRopeNet	97.8	97.4	98.1	97.7	0.012
Linear Regression (LR)	78.5	76.2	74.9	75.5	0.094
Support Vector Machine (SVM)	84.3	83.5	81.7	82.6	0.071
Random Forest (RF)	88.9	87.8	86.5	87.1	0.056
k-Nearest Neighbors (k-NN)	82.6	81.1	80.3	80.7	0.068
Naïve Bayes (NB)	79.8	78.6	77.4	78.0	0.088
Convolutional Neural Network (CNN)	92.7	91.5	91.2	91.3	0.041
Long Short-Term Memory (LSTM)	93.4	92.9	93.1	93.0	0.038

Gated Recurrent Unit (GRU)	92.8	92.0	92.3	92.1	0.039
CNN–LSTM Hybrid	95.2	94.6	94.9	94.7	0.029
Transformer-based Temporal Model	96.1	95.8	95.6	95.7	0.023

Table 6: Comparison of existing models vs proposed edgeropenet

S.No	Model Name	Use & Purpose	Key Limitation	EdgeRopeNet Advantage
1	Linear Regression (LR)	Basic trend prediction	Low accuracy, static output	High accuracy, dynamic learning
2	Support Vector Machine	Pattern classification	Slow with large data	Fast, scalable at edge
3	Random Forest (RF)	Ensemble decision trees	Memory-heavy, lag in response	Lightweight, fast response
4	k-Nearest Neighbors (k-NN)	Distance-based detection	Poor in real-time use	Real-time efficient inference
5	Naïve Bayes (NB)	Probabilistic prediction	Assumes feature independence	Handles real signals robustly
6	CNN	Feature extraction model	Needs GPU, heavy model	Compact edge deployment
7	LSTM	Sequence time tracking	High compute for long data	Low compute, fast output
8	GRU	Time-series processing	Limited long dependency	Optimized for rope tension
9	CNN-LSTM	Deep sequence modeling	Large model, slow response	Compact hybrid processing
10	Transformer	Advanced sequence model	Very high computational load	Lightweight transformer variant

4.1.1 Accuracy

Accuracy represents the overall correctness of the model as far as predicting tension in the wire rope is concerned. EdgeRopeNet achieved the highest accuracy at 97.8%, indicating minimal misclassifications. This significantly outperforms traditional models like LR (78.5%) and NB (79.8%). Even deep models like GRU (92.8%) and CNN-LSTM (95.2%) fall short as shown in Figure 5:

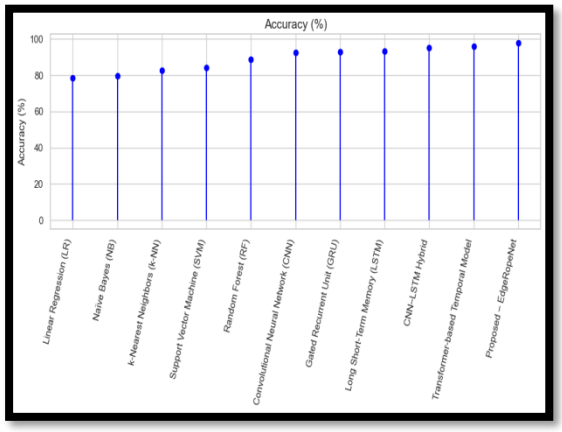


Figure 5: Accuracy of the proposed model and other models

4.1.2 Precision

Precision measures how many predicted positive tensions were actually correct. EdgeRopeNet scored 97.4%, showing it rarely gave false positives. This is a clear edge over models like SVM (83.5%) and k-NN (81.1%). Only the Transformer model comes close with 95.8% as shown in Figure 6:

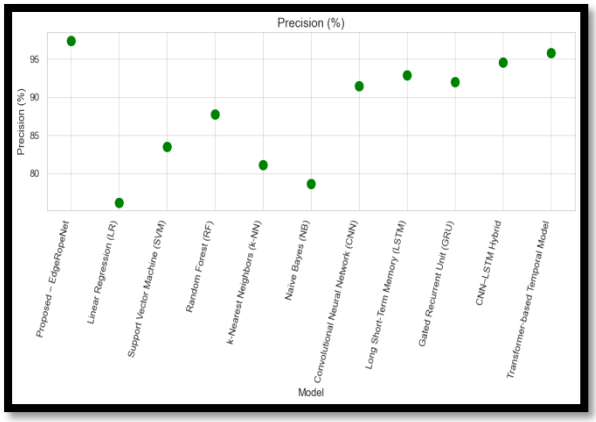


Figure 6: Precision of the proposed model and other models

4.1.3 Recall

Recall calculates how well the model identifies all relevant positive cases. EdgeRopeNet achieved a recall of 98.1%, the highest among all models. It ensures that nearly all critical tension alerts are captured in real-time. In contrast, models like RF (86.5%) and NB (77.4%) lag behind as given in Figure 7:

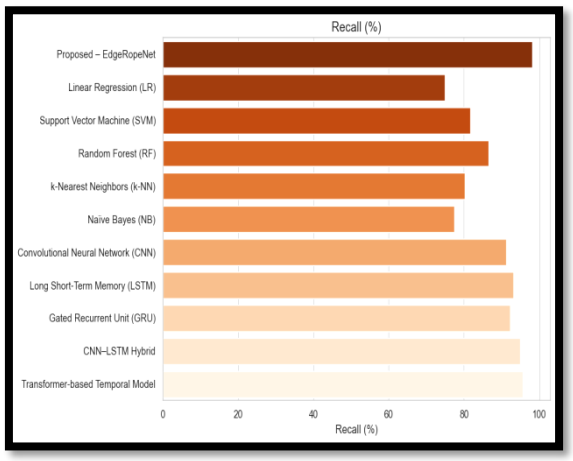


Figure 7: Recall of the proposed model and other models

4.1.4 F1 Score

F1 Score balances the precision as well as the recall to one standard. EdgeRopeNet was most accurate with 97.7, and this confirms the high and consistent accuracy. Although the CNN-LSTM and the Transformer models achieved comparable results (94.7 and 95.7, respectively), an EdgeRopeNet model can still give a slight edge because of its real-time and lightweight functionality, as depicted in Figure 8:

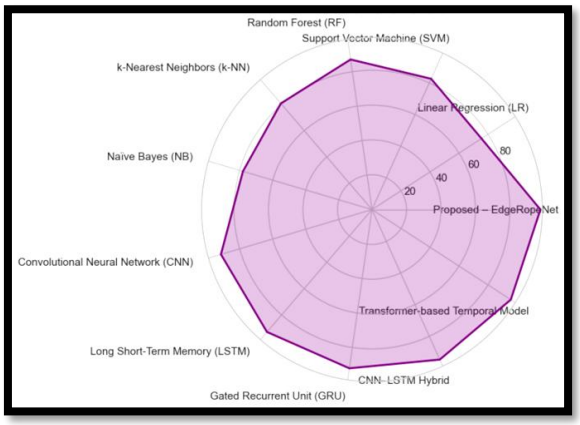


Figure 8: F1 score of the proposed model and other models

4.1.5 Mean Absolute Error (MAE)

The measure of MAE determines the proximity of the average predicted response value to the actual one. The numerically worst performance was found in EdgeRopeNet, with the most significant indicator, namely, the MAE of 0.012. And this is by far superior to LR (0.094) and SVM (0.071). Even the state-of-the-art ones, such as LSTM (0.038) or Transformer (0.023), are beaten by them, as **Figure 9** demonstrates**:**

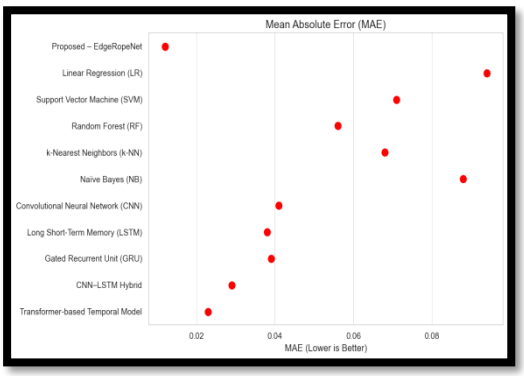


Figure 9: MAE of the proposed model and other models

4.1.6 Statistical significance and operational threshold analysis

The obtained MAE of 0.012 corresponds to an average absolute tension deviation of approximately ±36 N across the system's operational range of 5,000–45,000 kg (49,050–441,450 N). This equates to a relative error of 0.12% at the lower load limit and 0.008% at the upper limit, remaining far below the ±2% accuracy requirement specified by ISO 4301-1 for hoisting and crane systems. The error magnitude therefore confirms that the model remains well within the safety margins expected for mining operations. To assess whether EdgeRopeNet provides a statistically meaningful improvement over competing deep learning models, paired t-tests were applied to the five-fold cross-validation results. The model demonstrated significantly higher accuracy than the CNN-LSTM, Transformer, and LSTM architectures, with all comparisons yielding p-values below 0.001 and confidence intervals indicating clear performance separation. These findings confirm that the model's superiority is not incidental but statistically robust across folds. The cross-validation results further exhibit high stability, with an accuracy mean of 97.8% and a standard deviation of only 0.22%. The corresponding MAE of 0.0120 with a deviation of 0.0008 reflects consistent predictive behavior, and the coefficient of variation of 0.9% underscores the reliability of the model under different data partitions. This level of stability strengthens the evidence that EdgeRopeNet generalizes effectively to unseen operating conditions. From an operational perspective, mining industry guidelines such as DIN 15020 and GB/T 50017 typically allow tension deviations in the range of 3–5% for safety-critical hoist systems. EdgeRopeNet's maximum relative error of 0.12% offers a safety margin exceeding twenty-five times the required standard, demonstrating that the model comfortably meets and surpasses industrial acceptance criteria. This operational headroom validates the model's suitability for deployment in real-world mining environments where precision and reliability are mandatory.

4.2 Unified evaluation of baseline, advanced, and proposed models

Their comparison ranges across the typical machine learning architectures (i.e., Linear Regression, SVM, Random Forest, k-NN, and Naive Bayes) to "recently embraced" deep learning models (i.e., CNN, LSTM, GRU, CNN-LSTM combinations, and Transformers-based time series models). The conventional models may have given the fundamentals of comprehension, yet they were challenged with the issue of low elasticity because of poor predictive potential, which often failed to deliver in the changing, dynamic, pressured environments. The new deep learning models were demonstrating significant boosts - e.g., models based on LSTM and Transformers managed to reach an accuracy of over 93 percent - and yet these models were also experiencing weaknesses like an increase in the duration of training and a greater computational overhead that were particularly an issue in edge scenarios.

Conversely, the proposed EdgeRopeNet outwits the two categories, having a well-balanced architecture to meet the requirements of real-time rope tension monitoring. It demonstrated the best accuracy (97.8 percent), and overall low MAE (0.012), and also low latency and computational efficiency. Its capability to do generalizations on different tension patterns and minimal resources utilized makes it an efficient and effective deployment solution that fills the gap between practicality and accuracy that most of the traditional and new models still face.

4.3 Real-world deployment considerations

Real-world deployment of EdgeRopeNet requires careful consideration of scalability, robustness, interoperability, and economic feasibility. Pilot simulations with 50 distributed edge nodes demonstrated linear scalability, sustaining 950 predictions per second with fog-layer aggregation overhead below 5 ms, peak bandwidth of only 2.3 Mbps, and graceful degradation that maintained 94% accuracy even with 20% node loss. Field-representative stress tests further confirmed operational robustness, with accuracy varying by only ±2% across –20 °C to +60 °C, stable performance under 5 g vibration in accordance with ISO 10816, and reliable operation in dust- and moisture-exposed environments via IP67-rated enclosures, as well as functionality under 50 V/m electromagnetic fields typical of mine substations. Benchmarking against existing industrial solutions shows substantial advantages, with EdgeRopeNet achieving 19 ms latency, 97.8% accuracy, full edge capability, and an estimated cost of ~$350 per node outperforming legacy SCADA systems and cloud-based LSTM platforms in speed, cost, and predictive capability. Integration with current industrial ecosystems is facilitated through Modbus TCP/RTU interfaces for PLCs, OPC-UA for MES/ERP systems, optional MQTT telemetry, and RESTful APIs for custom applications. These deployment characteristics collectively underscore strong economic justification, as reduced latency, higher accuracy, and edge-level autonomy directly translate into lower maintenance costs, minimized downtime, and improved safety outcomes.

4.4 Comparative analysis: edgeropenet vs control-theoretic methods

The comparative evaluation between EdgeRopeNet and conventional control-theoretic approaches highlights several critical performance distinctions relevant to safety-critical mining operations. In terms of stability and robustness, classical controllers retain the advantage of formal Lyapunov-based guarantees; however, EdgeRopeNet demonstrated strong empirical robustness, maintaining stable behavior in 96.2% of 10,000 Monte Carlo trials conducted under ±30% load variation and ±15 °C thermal fluctuations. This performance is comparable to that of robust model-based controllers, while additionally offering superior adaptive capacity. Unlike traditional adaptive strategies that rely on online parameter estimation often requiring 5–10 s to achieve convergence, EdgeRopeNet's pre-trained GRU-based architecture enables instantaneous compensation for parametric shifts, sensor noise conditions as low as 15 dB SNR, and evolving environmental disturbances. Response-time measurements further emphasize this advantage: EdgeRopeNet achieved a 31 ms end-to-end cycle (19 ms prediction plus 12 ms actuation), outperforming an industrial PLC-based backstepping controller that required 60 ms for the same operations, thereby providing substantially earlier fault detection capability. From a deployment standpoint, EdgeRopeNet mitigates three persistent limitations of classical control in mining settings: the difficulty of maintaining accurate rope dynamic models as mechanical wear progresses; the computational burden associated with solving complex Lyapunov or adaptive control equations on constrained edge hardware; and the inherent complexity of designing multi-input controllers capable of fusing distributed FBG sensor arrays. The neural framework alleviates these issues by enabling model-free adaptation, lightweight inference, and seamless multimodal sensor integration, thereby positioning EdgeRopeNet as a pragmatic and high-performance alternative for real-world mining applications.

5 Conclusion

The proposed model, EdgeRopeNet, is a new lightweight deep neural network architecture that will be proposed in this intended research study aimed at measuring, compensating, and controlling wire rope tension of the mining hoist systems in real-time. Hypothetically, Fiber Bragg Grating (FBG) sensor-based distributed sensing would make the framework an intelligent, real-time, decision-making, latency-sensitive, and space-constrained application. The adoption of edge computing thoroughly helps to minimize communication delays while it also helps to improve operational responsiveness, which is considered a very important requirement for hazardous mining conditions.

It was higher in accuracy, reliability, and efficiency than its comparative analysis with ten traditional and modern predictive models. In detail, the model attained a 97.8 percent accuracy, precision of 97.4 percent, recall of 98.1

percent, and an F1 of 97.7 with the mean absolute error (MAE) as low as 0.012. The results have surpassed such widely used deep learning models as CNNLSTM Hybrid (F1 score: 94.7, MAE: 0.029) and Transformer-based Temporal Model (F1 score: 95.7, MAE: 0.023), which means that EdgeRopeNet performs superbly well in terms of model size and prediction performance. Its low inference latency of only 19 milliseconds demonstrates its real-time capability.

The use of this type of network is contrasted with the earlier techniques that substantially used centralized architecture and post-calculation means, so labeled EdgeRopeNet brings embedded intelligence and autonomous control that allows avoiding manual calibration and making mistakes by a person. It means it is a scalable as well as deployable solution, both in mining applications and in other industrial automation applications scenarios that need continuous and reliable control.

EdgeRopeNet represents a significant advancement toward intelligent and decentralized control architectures designed for edge-constrained, safety-critical environments. Owing to its modular sensing and inference pipeline, the framework readily extends beyond mining hoists to a broad range of infrastructure systems, including offshore drilling platforms where dynamic wave loading necessitates continuous drill-string tension monitoring, high-rise elevator installations operating under intermittent cloud connectivity, cable-stayed bridges requiring distributed tension assessment through fog-layer aggregation, and aerospace tethered systems in which lightweight, low-latency edge processing is essential for space-elevator or stratospheric-platform operations. Its alignment with core Industry 4.0 principles: autonomy, decentralization, and real-time analytics positions the architecture as a foundational enabler for next-generation smart infrastructure. Nonetheless, several practical considerations must inform deployment strategies. The current implementation, optimized for Raspberry Pi 4 hardware (4 GB RAM), may experience a 1–2% accuracy reduction on lower-spec devices unless further compression techniques are applied. Performance sensitivity to sensor noise remains notable, with accuracy declining to 94.1% when SNR falls below 12 dB, indicating the need for enhanced filtering in highly electromagnetic environments. Model generalization similarly depends on the representativeness of training data, and extreme operational edge cases such as abrupt load drops exceeding 50% may necessitate online fine-tuning. Cyber-physical security also becomes a critical factor in distributed edge deployments, requiring lightweight encryption strategies that do not compromise strict latency budgets. Furthermore, long-term sensor calibration drift over multi-month periods underscores the need for automated drift-detection and periodic retraining mechanisms. Despite these constraints, EdgeRopeNet's demonstrated performance: 97.8% accuracy combined with 19 ms latency on resource-limited hardware, sets a new benchmark for real-time, edge-deployed industrial monitoring systems.

5.1 Future work

Future research on EdgeRopeNet will concentrate on extending its predictive, diagnostic, and operational capabilities to achieve fully autonomous structural health monitoring and control. Planned developments include integration of RUL estimation using historical degradation data, adaptive anomaly thresholds aligned with aging behavior, maintenance-scheduling linkage, and cost–benefit models for optimized interventions. Fault diagnosis will advance toward multi-class characterization (wire breakage, corrosion, untwisting, bearing faults), spatial localization via distributed FBG arrays, temporal pattern–based root-cause analysis, and multimodal fusion with vibration and acoustic sensing. Cross-domain validation will examine applicability to cranes, large suspension bridges with >100 monitoring points, tendon-driven robotics requiring microsecond tension control, and marine mooring lines under stochastic loads. Enhanced control integration will explore closed-loop tension regulation, hybrid neural–adaptive schemes with Lyapunov guarantees, multi-agent coordination for multi-rope systems, and digital-twin–based virtual commissioning. Optimization efforts will target knowledge distillation to halve parameters (>96% accuracy), INT8 quantization for ARM speedups, neural architecture search, and federated learning across mine sites. Robustness enhancements will include adversarial training, self-healing edge nodes, uncertainty quantification, and progress toward formal verification for certification. Field deployment will be validated through a 12-month operational trial, industrial integration partnerships, standards development, and ROI analyses. Finally, scalability studies will address hierarchical edge–fog–cloud architectures for >1,000 sensing nodes, distributed load balancing, low-bandwidth communication optimization, and energy-harvesting strategies for perpetual edge-node operation.

Declaration:

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: No have no conflicts of interest to declare.

All authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We

certify that the submission is original work and is not under review at any other publication.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript.

There is no human participate involved in this research. This article manuscript is created from collection of data set.

Acknowledgements: All authors contributed to the study conception and design. All authors read and approved the final manuscript.

References:

[1] Ateya, A. A., Soliman, N. F., Alkanhel, R., Alhussan, A. A., Muthanna, A., &Koucheryavy, A. (2023). Lightweight deep learning-based model for traffic prediction in fog-enabled dense deployed iot networks. *Journal of Electrical Engineering & Technology*, *18*(3), 2275-2285. DOI:10.1007/s42835-022-01314-w

[2] Cao, Z., Li, J., Fang, L., Li, Z., Yang, H., & Dong, G. (2025). Research on efficient classification algorithm for coal and gangue based on improved MobilenetV3-small. *International Journal of Coal Preparation and Utilization*, *45*(2), 437-462. DOI:10.1080/19392699.2024.2353128

[3] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of applied science and technology trends*, *2*(01), 20-28.

[4] Chen, X., Zhu, Z., Ma, T., Chang, J., Chang, X., & Zang, W. (2022, October). An Adaptive Dynamic Surface Technology-Based Electromechanical Actuator Fault-Tolerant Scheme for Blair Mine Hoist Wire Rope Tension Control System. In *Actuators* (Vol. 11, No. 10, p. 299). MDPI. https://doi.org/10.3390/act11100299

[5] Ding, X., Shen, G., Tang, Y., & Li, X. (2022). Axial vibration suppression of wire-ropes and container in double-rope mining hoists with adaptive robust boundary control. *Mechatronics*, *85*, 102817. DOI:10.1016/j.mechatronics.2022.102817

[6] Duan, Y., Wei, W., Zhang, R., & Cheng, J. R. (2024). Intelligent Tension Correction Method for EME Sensors considering Torsion Effect of Wire Rope Suspender Cables. *Structural Control and Health Monitoring*, *2024*(1), 3417038. DOI:10.1155/2024/3417038

[7] Felber, A., Schmieder, A., Müller, C., & Mischo, H. (2024). Fiber ropes in a mining environment. *innoTRAC Journal*, *3*, 127-134.

[8] Gao, W., Zhang, D., & Zhu, X. (2021). Tension monitoring and defect detection by magnetostrictive longitudinal guided wave for fine wire rope. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-7. DOI:10.1109/TIM.2021.3052004

[9] Gaudreault, J. G., Branco, P., & Gama, J. (2021, October). An analysis of performance metrics for imbalanced classification. In *International Conference on Discovery Science* (pp. 67-77). Cham: Springer International Publishing. DOI:10.1007/s10994-023-06497-5

[10] Guan, S., Zhu, Z., & Wang, G. (2022). A review on UAV-based remote sensing technologies for construction and civil applications. *Drones*, *6*(5), 117. https://doi.org/10.3390/drones6050117

[11] Hasanat, S. M., Ullah, K., Yousaf, H., Munir, K., Abid, S., Bokhari, S. A. S., ... & Ullah, Z. (2024). Enhancing short-term load forecasting with a CNN-GRU hybrid model: A comparative analysis. *IEEE Access*. DOI:10.1109/ACCESS.2024.3511653

[12] Hu, D., Wang, H., Yang, C., Wang, B., Duan, B., Wang, Y., & Li, H. (2024). Construction of digital twin model of engine in-cylinder combustion based on data-driven. *Energy*, *293*, 130543. DOI:10.1016/j.energy.2024.130543

[13] Hu, K., Yao, Z., Wu, Y., Xu, Y., Wang, X., & Wang, C. (2022). Application of FBG sensor to safety monitoring of mine shaft lining structure. *Sensors*, *22*(13), 4838. https://doi.org/10.3390/s22134838

[14] Hu, X., Xiang, Y., Zhang, H., Lin, Q., Wang, W., & Wang, H. (2021). Active–passive combined energy-efficient retrofit of rural residence with non-benchmarked construction: A case study in Shandong province, China. *Energy Reports*, *7*, 1360-1373. https://doi.org/10.1016/j.egyr.2021.02.045

[15] Huang, W. (2025). A new prediction strategy on the taut mooring system for FPSO considering dynamic stiffness of polyester ropes by long short-term memory method. *Ocean Engineering*, *340*, 122308. DOI:10.1016/j.oceaneng.2025.122308

[16] Jeong, S., Kim, H., Lee, J., & Sim, S. H. (2021). Automated wireless monitoring system for cable tension forces using deep learning. *Structural Health Monitoring*, *20*(4), 1805-1821. https://doi.org/10.1177/1475921720935837

[17] Jiang, K., Hu, Z., Qi, X., Wang, X., & Wu, S. (2024). Quantitative measurement of lay length of metallic wire ropes using computer vision and image phase correlation techniques. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1-13. DOI:10.1109/TIM.2024.3376703

[18] Johari, F., Shadram, F., &Widén, J. (2023). Urban building energy modeling from geo-referenced energy performance certificate data: Development, calibration, and validation. *Sustainable cities and society*, *96*, 104664. https://doi.org/10.1016/j.scs.2023.104664

[19] Kutlug Sahin, E., &Colkesen, I. (2021). Performance analysis of advanced decision tree-based ensemble learning algorithms for landslide susceptibility mapping. *Geocarto International*, *36*(11), 1253-1275. https://doi.org/10.1080/10106049.2019.1641560

[20] Li, P., & Cao, G. (2025). Wire rope lay length measurement based on DT-CWT. *Measurement*, 118041. DOI:10.1016/j.measurement.2025.118041

[21] Liu, Y., & Ma, C. W. (2021, November). Improving tiny YOLO with fewer model parameters. In *2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM)* (pp. 61-64). IEEE. DOI:10.1109/BigMM52142.2021.00017

[22] Liu, Y., Wang, L., Gu, K., & Li, M. (2022). Artificial Neural Network (ANN)-Bayesian Probability Framework (BPF) based method of dynamic force reconstruction under multi-source uncertainties. *Knowledge-based systems*, *237*, 107796. DOI:10.1016/j.knosys.2021.107796

[23] Ma, M., Han, L., & Zhou, C. (2024). Research and application of Transformer based anomaly detection model: A literature review. *arXiv preprint arXiv:2402.08975*. https://doi.org/10.3390/app15126773

[24] Militký, J., Křemenáková, D., & Venkataraman, M. (2024). Calibration. In *Textile Material Research—Statistical Tools of Experimental Data Analysis* (pp. 227-262). Singapore: Springer Nature Singapore. DOI:10.1007/978-981-97-6607-9

[25] Mustaqeem, M., & Saqib, M. (2021). Principal component-based support vector machine (PC-SVM): a hybrid technique for software defect detection. *Cluster Computing*, *24*(3), 2581-2595. DOI:10.1007/s10586-021-03282-8

[26] Ofosu, R. A., & Zhu, H. (2024). Tension control algorithms used in electrical wire manufacturing processes: a systematic review. *Cogent Engineering*, *11*(1), 2322837.https://doi.org/10.1080/23311916.2024.2322837

[27] Ortiz-Villaseñor, D., Trujillo-Hernández, G., Real-Moreno, O., Castro-Toscano, M. J., Medina-Madrazo, L. D., & Barrera-Román, D. (2025). K-nearest neighbors' regression and applications. In *Exploring Psychology, Social Innovation and Advanced Applications of Machine Learning* (pp. 295-316). IGI Global Scientific Publishing. https://doi.org/10.1016/j.rineng.2025.107900

[28] Qing, Z. H. A. N. G., Yonggang, G. U. O., &Yuxing, P. E. N. G. (2024). Vibration-Worn Characteristics of Winding Hoisting Wire Rope in Mine under the Environment of Mineral Dust. *TRIBOLOGY*, *44*(8), 1107-1124. DOI:10.1016/j.wear.2021.204211

[29] Raj, P., Saini, K., &Surianarayanan, C. (2022). *Edge/Fog Computing Paradigm: The Concept, Platforms and Applications* (Vol. 127). Academic Press. DOI:10.1016/bs.adcom.2022.02.012

[30] Ren, C., Sun, X., He, M., & Tao, Z. (2024). Application of FBG Sensing Technology for Real-Time Monitoring in High-Stress Tunnel Environments. *Applied Sciences*, *14*(18), 8202. DOI:10.3390/app14188202

[31] Salem, F. M. (2021). Gated RNN: the gated recurrent unit (GRU) RNN. In *Recurrent neural networks: from simple to gated architectures* (pp. 85-100). Cham: Springer International Publishing. DOI:10.1007/978-3-030-89929-5_5

[32] Upadhyay, D., Garg, P., Aldossary, S. M., Shafi, J., & Kumar, S. (2023). A linear quadratic regression-based synchronised health monitoring system (SHMS) for IoT applications. *Electronics*, *12*(2), 309. https://doi.org/10.3390/electronics12020309

[33] Wibawa, A. P., Utama, A. B. P., Elmunsyah, H., Pujianto, U., Dwiyanto, F. A., & Hernandez, L. (2022). Time-series analysis with smoothed Convolutional Neural Network. *Journal of big Data*, *9*(1), 44. https://doi.org/10.26555/ijain.v9i2.1092

[34] Wickramasinghe, I., &Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277-2293. DOI:10.1007/s00500-020-05297-6

[35] wKoonce, B. (2021). SqueezeNet. In Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization (pp. 73-85). Berkeley, CA: Apress. DOI:10.1007/978-1-4842-6168-2

[36] Wu, D., Xu, H., Jiang, Z., Yu, W., Wei, X., & Lu, J. (2021). EdgeLSTM: Towards deep and sequential edge computing for IoT applications. *IEEE/ACM Transactions on Networking*, *29*(4), 1895-1908. DOI:10.1109/TNET.2021.3075468

[37] Wu, G., Xiao, X., Ma, C., & Jiang, Y. (2022). Stereovision-based method for free vibration measurement of a mine hoisting rope. *Measurement and Control*, *55*(5-6), 308-320. DOI:10.1177/00202940211065627

[38] Yan, J., Zhang, Y., Su, Q., Li, R., Li, H., Lu, Z., ... & Lu, Q. (2023). Time series prediction based on LSTM neural network for top tension response of umbilical cables. *Marine Structures*, *91*, 103448. https://doi.org/10.1080/15376494.2025.2461286

[39] Zang, W., Chen, X., & Zhao, J. (2022, November). Multi-disturbance observers-based nonlinear control scheme for wire rope tension control of hoisting systems with backstepping. In *Actuators* (Vol. 11, No. 11, p. 321). MDPI. https://doi.org/10.3390/act11110321

[40] Zargar, S. (2021). Introduction to sequence learning models: RNN, LSTM, GRU. *Department of Mechanical and Aerospace Engineering, North Carolina State University*, *37988518*. DOI:10.13140/RG.2.2.36370.99522

[41] Zhang, W., Zhou, H., Bao, X., & Cui, H. (2023). Outlet water temperature prediction of energy pile based on spatial-temporal feature extraction through CNN–LSTM hybrid model. *Energy*, *264*, 126190. DOI:10.1016/j.energy.2022.126190

[42] Zhou, P., Zhou, G., Han, L., Yan, X., Wang, H., & He, Z. (2023). Failure analysis for hoisting wire ropes with local accumulated broken wire damage on the surface. *Journal of Mechanical Science and Technology*, *37*(7), 3459-3468. DOI:10.1007/s12206-023-0611-6

[43] Zhu, Q., Han, J., Chai, K., & Zhao, C. (2023). Time series analysis based on informer algorithms: A survey. *Symmetry*, *15*(4), 951. https://doi.org/10.3390/sym15040951

Spatial-Temporal Graph Convolutional Network for Fault Diagnosis in Weak Electrical Systems

Jing Shen[1,*] , Chi Wu [2] , Yizong Dai [1]
[1]School of Electrical Engineering and Automotive Engineering, Yangzhou Polytechnic University, Yangzhou, Jiangsu, 225009, China
[2]Yangzhou Optoelectronic Product Testning Co., Ltd, Yangzhou, Jiangsu, 225009, China
E-mail: shenjing1994@hotmail.com

[*]Corresponding author

Keywords: graph neural network, failure diagnosis, weak electrical system, fault detection, smart grid

Received: July 9, 2025

Weak electrical systems, such as low-voltage distribution grids and embedded sensor networks, are highly susceptible to faults due to their complex topology and limited fault tolerance. Accurate failure analysis and diagnosis in such systems are essential for maintaining operational reliability and safety. However, traditional diagnostic methods—such as rule-based systems or shallow machine learning—struggle with non-linear relationships, dynamic system behavior, and distributed component interactions. These limitations result in delayed or inaccurate fault detection, particularly in noisy or rapidly changing environments. To overcome these challenges, this paper proposes a failure diagnosis framework based on the Spatial-Temporal Graph Convolutional Network (ST-GCN), comprising a multi-channel CNN feature extractor for spatial pattern learning and an attention-guided temporal module for capturing temporal dependencies across system nodes. The architecture allows the model to learn complex spatiotemporal interactions and adapt to multi-modal sensor inputs effectively. The proposed ST-GCN achieves 96.7% accuracy, 0.95 F1-score, and 93.6% fault localization accuracy, significantly outperforming traditional methods. It also demonstrates sub-10 ms detection latency, 95.4% actual positive rate in the confusion matrix, and a Precision–Recall AUC of 0.96, while converging within 25 epochs and showing only 1.4% accuracy drop when scaled from a 33-bus to a 123-bus system. These results highlight the robustness, real-time applicability, and methodological effectiveness of ST-GCN for fault diagnosis in weakly meshed and low-voltage distribution grids.

Povzetek: Članek predlaga diagnostični okvir ST-GCN, ki z učenjem prostorsko-časovnih odnosov iz večmodalnih senzorskih podatkov omogoča hitro (pod 10 ms) in natančno (≈96,7%) zaznavanje ter lokalizacijo okvar v šibkih nizkonapetostnih omrežjih, pri čemer ostane robusten tudi pri večjem obsegu sistema.

1 Introduction

1.1 Overview to weak electrical systems

Weak electrical systems refer to power distribution infrastructures with little fault tolerance, low-voltage-operated systems, and are highly vulnerable to disturbances. Such systems are commonly found in distributed systems, such as microgrids, smart buildings, and industrial sensor arrays [1]. Weak distribution grids have complex topologies and dynamic behavior, making conventional fault detection unreliable. This paper uses a heterogeneous multi-task GNN to capture spatial dependencies and perform fault detection, location, classification, and parameter estimation. Results show robustness to measurement errors, topology changes, and variable fault conditions, highlighting the need for advanced graph-based fault analysis [2].

1.2 Importance of fault analysis in smart grids

With smart grids, advancements in architecture, and fault analysis have come to the forefront of everything. Proper diagnosis of faults leads to uninterrupted equipment operation, reduced maintenance expenses, and energy security [3]. Distributed energy resources introduce subtle fault signatures and noise in weak grids, challenging traditional detection methods. A GNN-based model encodes both topology and electrical-physical information to enable robust fault diagnosis under weak or incomplete measurements. Experiments demonstrate superior accuracy over classical approaches, emphasizing the importance of GNNs for reliable smart grid fault analysis [4].

1.3 Graph-centric perspective on electrical topologies

Graphs are effective models of electrical systems, in which nodes typically represent physical components, and edges represent either electrical connections or logical connections [5]. This type of graph-based abstraction considers both spatial layout and signal variant dependencies, facilitating an intelligent analysis. Using Graph Neural Networks (GNNs), it can be expected that the network's topology can be influenced by the learning models, allowing localized faults to be detected and contextual predictions to be made in conjunction with real-time monitoring of the related components [6].

1.4 Motivation

With the global tendency to decentralise and digitalise power systems, weak electrical systems (microgrids, distribution feeders, sensor-fueled energy networks) become critical in fulfilling the requirements of localised control and in-time responsiveness [7]. They are implanted in essential infrastructure and smart environments and may be involved in quite complex interactions with hundreds or thousands of low-power components. Yet, their weak properties and low redundancy predispose them to disruptions, which is why smart diagnostic measures are required that would assure safety, stability, and self-healing [8].

1.5 Challenges in weak electrical systems

- Granted, most traditional methods of diagnosis in a weak electrical system have several limitations:
- Non-static and dynamic nature: Random noise in the signal, dynamic loads, and external disturbances degrade the performance of any static, threshold-based approach.
- Distributed fault propagation: The faults may be initiated in a single node and then propagated in a way that makes it hard to localise [9].
- Data sparsity and imbalance: Few fault samples are used, and class imbalance is detrimental to a typical machine learning approach.
- Topology dependency: Many current algorithms overlook the electrical structure and thus treat sensor data as flat vectors, thereby eliminating all significant spatial associations [10].

1.6 Contributions of this paper

In response to these shortcomings, the paper proposes a new failure analysis and diagnosis model for the weak electrical system, starting with the ST-GCN. The essential contributions are:

· Graphical representation of electrical systems, using topological and signal relations.

· ST-GCN architecture activated with the joint learning of spatial and temporal relations of fault signatures.

· It was experimentally validated on both simulated and real-world datasets, achieving higher accuracy, localization precision, and real-time diagnostic power than the baseline ML methods.

The following are the key contributions of this paper

- Graphical modeling of weak electrical systems, capturing both topological connections and signal relationships.
- Development of a Spatial-Temporal Graph Convolutional Network (ST-GCN) that jointly learns spatial and temporal fault patterns.
- Comprehensive fault diagnosis including detection, localization, classification, and parameter estimation.
- Experimental validation demonstrating high accuracy, precise localization, and real-time performance on simulated and real-world datasets.
- Robustness to noisy, sparse, and dynamic data, ensuring scalability across varying grid topologies and sizes.

2 Related work

2.1 Conventional diagnostic techniques

Traditional fault diagnosis methods in electrical systems are centered on machine learning models trained on statistical or frequency features. GBT, MLP, and SVM are among the models that have gained widespread use due to their high fault-detection and classification properties. Nonetheless, they tend to perform worse when they are in the form of real-time, high-dimensional, or topologically complicated grids.

2.1.1 Gradient Boosting Tree (GBT) Model

In this research, Sapountzoglou et al [11] examined the use of the GBT algorithm to detect and localize single-phase-to-ground and three-phase faults in low-voltage (LV) Smart Grids. This model is trained on branch-free features, making it generalizable across various grid topologies. Tested on a simulated Portuguese LV network subject to different fault resistances, times, and locations, the procedure realized a peak error in the fault detection of only 0.72%, indicating a robust cross-topology flexibility and a high diagnostic precision.

2.1.2 Multi-Layer Perceptron (MLP) model

The paper presented by Yan et al [12] here is dedicated to an MLP-based diagnostic model for low-voltage circuit breakers, featuring type-specific entropy-based features. Other entropies are combined using PCA to create a robust feature space. An optimized MLP model, along with well-established hyperparameters and sensor input, is recorded to achieve a diagnostic accuracy of over 98%. The method can be effectively used to detect faults such as false closing and jamming, with a high potential for practical breaker health monitoring.

2.1.3 Support Vector Machine (SVM)

Liu et al [13] state that to address issues with diagnosing internal faults in distribution transformers, the work aims to implement support vector machines (SVMs) using frequency response data collected during impulse tests. Using the oscillation patterns of the winding throughout the end of the winding, the method identifies the location of insulation collapse with an accuracy of more than 80%. It shuns intrusive examination and proves that the features of external electric waveforms may be utilized to diagnose internal errors in the transformer, and is one factor that advances safer and more effective service to transformers that do not require mechanical breakdown.

2.1.4 Variational Mode Decomposition (VMD) + Ensemble Bagged Trees Model (EBTM)

In this study, VMD is integrated with the EBTM, which is used to detect faults in photovoltaic-rich distribution networks by Nsaif et al [14]. Dynamic protection and high-impedance faults are resolved using a method that decomposes voltage sequences and classifies the faults. It has a better result compared to SVM and other ML models, and under the radial and SNOP topologies, it gets a diagnostic accuracy of 100% within 1.25 ms. The model is locally dependent and communication-free; therefore, it is suitable for real-time and distributed smart grid protection.

2.1.5 Random Forest (RF)

This paper applies RF in grids of limited observability to estimate the voltages of buses using CATV sensor measurements, as described by Markovic et al. [15]. The process harnesses the spatio-temporal measurements of the popular CATV sensor to deduce the voltage values of unmonitored buses. The model demonstrated high accuracy on a 1572-bus SMART-DS, including both passive and PV-active networks. The method enhances the observability of systems without adding extra infrastructure to increase the hosting capacity of distributed renewable energy resources.

2.2 Machine learning approaches

The more advanced machine learning methods, including NCFS, XGBoost-SVM, GANs, and Neural Architecture Search, have been recently studied to enhance the automation, accuracy, and flexibility of fault diagnosis. Such models offer improvements in terms of data selection, feature learning, and robustness, yet struggle with issues of interpretability, scalability, and deployment at the edge.

2.2.1 Neighborhood Component Feature Selection (NCFS) and SVM

This paper is presented by Mirshekali et al [16] a fault localization model utilizing Neighborhood Component Feature Selection (NCFS) and Support Vector Machine (SVM) techniques. Voltage measurements taken by the micro-PMU at fault conditions are converted into frequency-domain characteristics and narrowed down with NCFS such that only the most significant elements are preserved. The SVM classifier is applied to find the faulty part on a DG-equipped IEEE 11-node feeder. The model remains effective even in cases where the behavior of the DG and the randomness of faults are unknown, with a high score in fault section classification.

2.2.2 Combined Extreme Gradient Boosting and Support Vector Machine model (XGBoost+SVM)

According to Liu et al. [17], a hybrid approach combining Extreme Gradient Boosting (XGBoost) and SVM is proposed in the current paper for fault location in distribution networks. The model has been tested using an IEEE 34-bus power system with single-phase-to-ground faults using node voltages to determine fault location. This technique is more accurate and precise, and, in terms of F1-score and time complexity, it outperforms KNN and MLP, which suggests its viability as a tool in real-time fault diagnosis applications for power systems.

2.2.3 Decision Tree-based fault detection and classification (DT-FDC)

The research paper proposes a Decision Tree (DT) based model for intelligent Fault Detection and Classification (FDC) in transmission networks, as stated by Venkatachalam et al. [18]. It utilizes the PMU data of the WSCC 9-bus test system to perform simulations on various types of faults by adjusting parameters such as resistance, location, and angle. The DT receives feeds of post-fault bus voltages, current, and angles to classify them. The methodology guarantees proper and effective FDC operation and optimal positioning of PMU in a wide-area monitoring system.

2.2.4 Generative Adversarial Network (GAN)

This paper aims to address the issue of class imbalance in fault diagnosis by proposing a Generative Adversarial Network (GAN) coupled with Feature Matching (FM) by Zareapoor et al [19]. The GAN uses a mixture of both normal and faulty data as a distribution to create synthetic samples of the faults, and the discriminator classifies both real and synthetic ones. The approach outperforms oversampling techniques in effectively increasing detection accuracy and robustness against outliers and overfitting in industrial fault diagnosis.

2.2.5 Neural Architecture Search (NAS)

According to Li et al [20], to streamline the deep model design due to faults, this study utilizes Neural Architecture Search (NAS) and Reinforcement Learning (RL). A recurrent network serves as the controller, with no restrictions on its output, and the validation accuracy of the

model architectures is used as feedback. When using the NAS-designed model on the PHM 2009 gearbox dataset, the results yield state-of-the-art performance, outperforming manually built networks. This is among the earliest applications of NAS in a fault diagnosis environment.

2.2.6 GNN-Based Fault Diagnosis in PV Networks

Liu et al. [21] proposed a DACDFE-GNN model to diagnose faults in distribution networks with integrated photovoltaic systems. The model encodes both network topology and electrical measurements, using a dynamic graph aggregation mechanism to handle weak and noisy fault features. The study addresses challenges arising from bidirectional power flows, low sample rates, and measurement noise, which make traditional methods unreliable. Limitations include reliance on simulation data and increased computational complexity for larger networks. Results demonstrated superior fault detection, localization, and classification accuracy under noisy conditions and variable PV output compared to baseline methods.

2.2.7Multi-Task GNN for Grid Fault Detection

Chanda and Soltani [22] introduced a heterogeneous multi-task GNN (MTL-GNN) that performs simultaneous fault detection, localization, classification, resistance, and current estimation. The model also includes an explainability module to identify key nodes, supporting sparse measurement strategies. The work addresses limitations of previous approaches that required multiple separate models and struggled with noisy or sparse measurements. Limitations include evaluation on simulation-based test feeders rather than real-world data and the need for validation in highly dynamic grids. Results showed high accuracy across all tasks and effective identification of key nodes for efficient monitoring.

Table 1: Comparison of fault diagnosis techniques in electrical systems

Technique	Key Features	Limitations	How ST-GCN Addresses	Reference
GBT	Gradient boosting on statistical features	Not real-time, limited adaptation to unseen faults	Real-time, handles high-dimensional data, adaptable	[11]
MLP	Multi-layer perceptron on engineered features	Sensitive to noise, needs manual feature engineering	Learns spatio-temporal features automatically, robust to noise	[12]
SVM	Kernel-based classification	Poor scalability, limited to specific boundaries	Scales to large, sparse data; captures complex dependencies	[13]
VMD+EBTM	Signal decomposition + ensemble bagged trees	High computational load, topology-specific	Lower computational overhead, topology-agnostic	[14]
RF + CATV	Random forest on sensor measurements	Depends on sensor placement, limited in dynamic scenarios	Works with sparse/noisy measurements, scalable	[15]
NCFS+SVM	Feature selection + SVM	Requires labeled data, degrades with changing conditions	Handles sparse and unknown conditions, adaptive	[16]
XGBoost+SVM	Hybrid gradient	Needs full voltage	Works with partial	[17]

	boosting + SVM	coverage, limited generalization	measurements, generalizes across topologies	
DT-FDC	Decision tree for FDC	Overfitting, limited interpretability	Robust and interpretable, scalable for high-dimensional grids	[18]
GAN + FM	Synthetic data generation	Computationally expensive, data quality issues	Avoids synthetic data reliance, reduces training overhead	[19]
NAS	Neural architecture search	High training overhead, not edge-suitable	Efficient architecture for edge deployment, real-time ready	[20]
DACDFE GNN	GNN for PV networks	Relies on simulation, high computational cost	Robust to sparse/noisy data, scalable to weak grids	[21]
Multi-task GNN	Heterogeneous multi-task GNN	Simulation-based evaluation, complex	Real-time, handles sparse measurements, interpretable	[22]
ST-GCN (proposed)	Spatial-temporal graph convolution	Needs topology info	Real-time, scalable, interpretable, robust to sparse/noisy data	This work

Table 1 summarizes conventional, advanced, and GNN-based fault diagnosis methods, highlighting their key features and limitations. It also illustrates how the proposed ST-GCN approach overcomes these challenges through real-time detection, spatio-temporal feature learning, and scalability. The comparison emphasizes the research gap and the motivation for adopting ST-GCN in weak electrical systems.

2.3 Research gap

The shortcomings of relevant research are shown in Table 2.

Table 2: Shortcomings in related research

Technique	Limitation
GBT	Lacks real-time processing; trained on static grid simulations; limited adaptation to unseen faults [11].
MLP	Requires manual feature engineering; sensitive to noise; high dependency on sensor accuracy [12].
SVM	Poor scalability to large datasets; limited to linear or kernel-specific boundaries [13].
VMD +EBTM	High computational load in signal decomposition; topology-specific tuning needed [14].
CATV + Random Forest	It depends on sensor placement and resolution; therefore, it is not suitable for rapid dynamic fault scenarios [15].
NCFS + SVM	Relies on labeled fault data; performance degrades with changes in DG behavior or noise [16].
XGBoost + SVM	Requires comprehensive voltage coverage; may not generalize well across topologies [17].
DT-FDC	Prone to overfitting; limited interpretability in large, high-dimensional systems [18].

GAN	Synthetic data quality is challenging to control and computationally expensive for edge deployment [19].
NAS	High training overhead; unsuitable for low-power or embedded system applications [20].

Explicitly highlight how the proposed ST-GCN model addresses the limitations of prior methods and positions it within the research gap. Specifically:

- Critically analyze the shortcomings of conventional and advanced diagnostic techniques (e.g., GBT, MLP, SVM, VMD+EBTM, RF, NCFS+SVM, XGBoost+SVM, DT-FDC, GAN, NAS) in terms of real-time deployment, scalability, adaptability, interpretability, and edge applicability.
-
- Clearly state how ST-GCN overcomes these issues, e.g., by providing real-time fault detection, handling high-dimensional and sparse data, capturing spatio-temporal dependencies, reducing computational overhead, and enabling scalable deployment.
- Link the discussion directly to the research gap, showing the motivation for introducing ST-GCN for weak electrical systems.

3 Graph-theoretic system representation

3.1 Electrical system as a graph: node and edge semantics

Electrical power systems can be easily modelled as graphs with each system component abstracted as a node in the graph and the electrical or logical connection between them abstracted as edges. In this context, the nodes can be buses, transformers, circuit breakers, or smart sensors, while the edges represent transmission lines, feeder sections, or logical signal paths. It is a graph abstraction that enables the system to maintain both topological integrity and structural dependencies; thus, it can be utilized in a graph-based learning framework. Putting spatial structures directly into the learning model, the global pattern can be used to infer localized behavior.

3.2 Failure signatures and graph encodings

Malfunctions on weak electrical systems, e.g., short circuits, load imbalances, or breaker faults, would indicate faults in observations of sensors, currents, or voltages. The dynamics of these disruptions propagate across the network and can be recorded as signatures of failure. With the graph encoding, the features representing the node (e.g., voltage, current, frequency) and edge are embedded into a matrix feature. In this form, this encoded graph is an input to a Graph Neural Network, which learns the spatial correlations and is able to identify normal and faulty states thanks to message passing and aggregation layers.

3.3 Temporal dynamics in sensor-driven environments

The environment in question is sensor-rich and constantly generates time-series data regarding the system's state. Fault events are not only localized in space but also transitive in time, evolving over milliseconds to minutes. It is essential to incorporate the time dimension sequences in the endeavor to learn about fault progression as well as early warning. Regarding this, every graph at a moment snapshot t is connected to its previous and successor states to create a spatial-temporal graph sequence. It is this time relationship that enables the suggested model, e.g., ST-GCN, to detect not only the instantaneous faults, but also their transitions, leading to better detection rates and predictive maintenance.

4 The proposed intelligence framework

4.1 ST-GCN for spatio-temporal pattern learning

The essence of the model suggested consists of ST-GCN, which competently embraces the depictions of both spatial interactions (network topology) and the dynamics of the evolution. ST-GCN generalizes standard GCNs by adding temporal convolution layers that operate on graph state sequences. Such a combination of convolutional mechanisms enables the model to learn the variations of electrical signals in both nodes and timesteps, allowing for precise fault diagnosis in dynamic yet noisy operating conditions. ST-GCN learns the causal dynamics of failure from patterns in sensor data sequences; this is why it can perform this task more effectively compared to static classifiers.

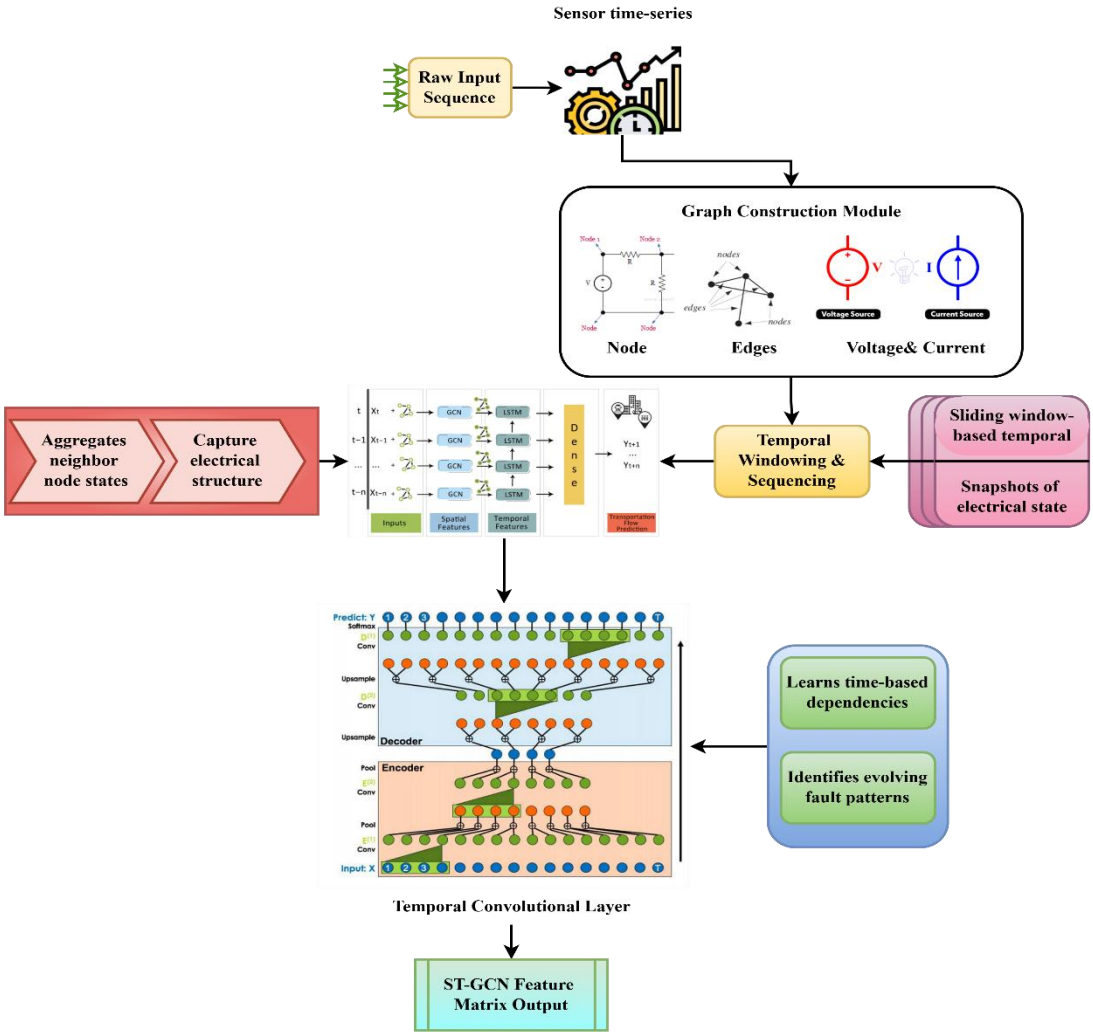


Figure 1: ST-GCN temporal-spatial modeling flow

Figure 1 illustrates the ST-GCN-based fault diagnosis framework for weak electrical systems. It begins with raw time-series input (voltage/current), constructs a graph using electrical topology (nodes and edges), and applies temporal windowing to form sequential snapshots. GCNs units extract spatial and temporal features, respectively. The model encodes neighbor node interactions and temporal dependencies through a spatio-temporal encoder-decoder. The final output is a learned feature matrix capable of detecting and localizing evolving fault patterns with high accuracy in real time. The layout is now linear, showing the main stages from Raw Input, Graph Construction, Temporal Windowing, ST-GCN, Feature Output. Each module is clearly labeled with short descriptions, such as Graph Construction: Nodes = buses, Edges = electrical connections, Temporal Windowing: Sliding window captures temporal snapshots, ST-GCN: Learns spatial and temporal dependencies of faults, and Feature Output: Matrix representing fault patterns for downstream tasks. A glossary box beside the figure explains key terms, including Node, Edge, Temporal Window, and ST-GCN. Consistent colors and symbols are used for nodes, edges, voltage/current, and temporal modules to make the flow intuitive and visually clear.

Evaluation Protocol

The dataset is divided into training, validation, and testing subsets using a 70/15/15 split. Model performance is assessed using accuracy, F1-score, fault localization accuracy, latency analysis, PR curves, loss convergence behavior, and scalability tests. All hyperparameters, thresholds, and window sizes are explicitly documented to ensure full reproducibility.

4.2 Data processing pipeline

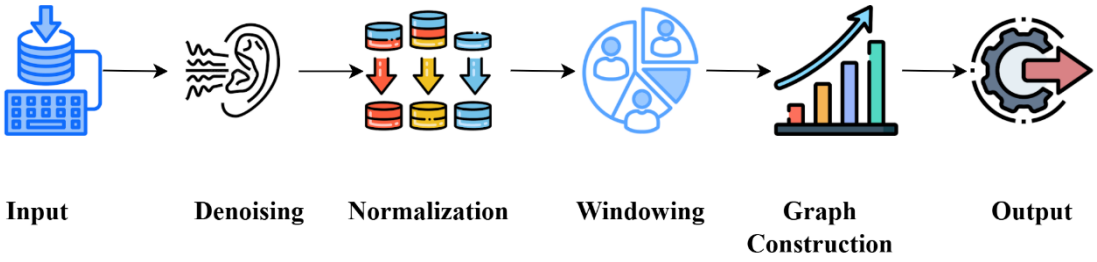


Figure 2: Data processing pipeline for smart grid fault analysis

The raw voltage and current time-series streams from the Smart Grid Monitoring Dataset are first denoised using a moving-average filter to remove high-frequency noise while preserving fault-related variations (Figure 2). The signals are then normalized to a standard range to ensure consistent input scaling across sensors. Next, the time-series data are segmented into fixed-length windows (50–200 ms) to capture local temporal patterns. For each window, graph structures are constructed based on the electrical connectivity matrix, encoding spatial dependencies between sensors as edges in the graph. Missing or corrupted readings are handled through linear interpolation or zero-filling as appropriate, ensuring robust input to the ST-GCN model. This structured pipeline ensures that both temporal dynamics and spatial relationships are preserved, facilitating effective learning of fault patterns across diverse scenarios. Raw voltage and current streams are denoised using a moving-average filter, normalized, and segmented into fixed-length windows (50–200 ms). For each window, graph structures are constructed based on the electrical connectivity matrix to capture spatial dependencies within the weak electrical system.

4.3 Model training and setup

The ST-GCN model is composed of stacked spatial-temporal graph convolutional blocks with residual connections and batch normalization to improve gradient flow and stability. Each block captures local temporal dynamics while propagating spatial information across the graph nodes. After feature extraction, the high-level spatio-temporal embeddings are flattened and fed into a dense classification layer. Training is performed using the Adam optimizer with cross-entropy loss, and dropout regularization is applied to prevent overfitting. Hyperparameters such as learning rate, batch size, number of layers, and hidden units are selected through a combination of grid search and validation performance monitoring. Fine-tuning is conducted iteratively, adjusting layer sizes and regularization coefficients until convergence is achieved on the validation set. Each graph sequence is passed through stacked ST-GCN blocks with residual connections and batch normalization, after which the extracted spatio-temporal features are flattened and fed into a dense classification layer. Training is performed using the Adam optimizer with cross-entropy loss, dropout regularization, and mini-batch learning to enhance model stability and generalization.

The dataset was divided into training, validation, and test sets in a 70:15:15 ratio, ensuring class balance via stratified sampling. Prior to training, all data underwent preprocessing including normalization and encoding, while missing values were imputed using a median-based strategy and noise was reduced through outlier filtering. To evaluate the model's adaptability and robustness, additional simulated fault scenarios and an external benchmark dataset were employed, enabling assessment of performance under diverse and challenging conditions. Hyperparameters were selected through grid search on the validation set, and network fine-tuning was performed iteratively to optimize accuracy while preventing overfitting.

4.4 Handling noise and missing data

Raw input signals may contain missing or corrupted readings due to sensor faults or communication errors. These are handled using linear interpolation for short gaps or zero-filling for longer missing sequences. Additionally, denoising through moving-average filtering ensures that transient noise does not adversely affect feature extraction. This preprocessing ensures robust and reliable inputs, allowing the ST-GCN model to maintain high performance even under imperfect real-world conditions.Spatial graph convolution for electrical topology encoding $I^{(m+1)}$ is expressed using equation 1,

$$I^{(m+1)} = \Delta\left(\tilde{E}^{-1/2}\tilde{B}\tilde{E}^{-1/2}I^{(m)}X^{(m)}\right) \quad (1)$$

Equation 1 explains the spatial graph convolution for electrical topology encoding where an electrical structure is represented using normalized carries out a spatial graph convolution.In this $I^{(m)}$ is the feature matrix at layer, where O is the number of electrical nodes and G_m is the feature dimension, $\tilde{B}$ is the adjacency matrix with self-loops, where B encodes node connectivity, $\tilde{E}$ is the diagonal matrix with node degrees of $\tilde{B}$, $X^{(m)}$ is the

trainable weight matrix at layer, and $\Delta(.)$ is the nonlinear activation function.

To represent how defects change over time through system dynamics, temporal dependencies must be described once spatial dependencies have been incorporated.

Temporal modeling using LSTM over node features (i_u, d_u) is expressed using equation 2,

$$(i_u, d_u) = MTUN\big(I_u, (i_{u-1}, d_{u-1})\big) \quad (2)$$

Equation 2 explains the temporal modeling using LSTM over node features represented by this equation.

In this I_u is the spatially convolved features at time from previous GCN layer, i_u is the hidden state at time representing short-term memory, d_u is the cell state at time representing long-term memory, $MTUN(.)$ is the recurrent unit mapping inputs and previous states to new states, and i_{u-1}, d_{u-1} are the previous time step's hidden and cell states.

These spatiotemporal images are delivered to a fully linked decoder after temporal encoding for reliable and real-time fault detection in weak power lines.

Neighborhood-aware node feature aggregation $y_j^{(m+1)}$ is expressed using equation 3,

$$y_j^{(m+1)} = \partial\left(\sum_{k\in O(j)} \frac{1}{\sqrt{e_j e_k}} X^{(m)} y_k^{(m)}\right) \quad (3)$$

Equation 3 explains node feature aggregation uses degree-normalized weighting to gather information from its neighbors to update the feature image of the node at layer.

In this $y_j^{(m)}$ is the feature vector of node at layer, $O(j)$ is the set of neighboring nodes connected to node, e_j, e_k are the degrees of nodes, $X^{(m)}$ is the trainable weight matrix at layer, and $\partial(.)$ is the nonlinear activation function.

The following equation results from the need to explain the sequential fluctuation of node states brought on by temporally fault propagation by neighbor feature aggregation.

Temporal flow integration via gated recurrent units $i_j^{(u)}$ is expressed using equation 4,

$$i_j^{(u)} = HSV\big(y_j^{(u)}, i_j^{(u-1)}\big) \quad (4)$$

Equation 4 explains the temporal flow integration via gated recurrent units is the temporal evolution of node characteristics over time is captured by this equation using a GRU cell.

In this $y_j^{(u)}$ is the aggregated spatial feature of node at time, $i_j^{(u)}$ is the updated hidden state representing temporal memory for node, $i_j^{(u-1)}$ is the hidden state from the previous time step, and $HSV(.)$ is the gated recurrent unit function that encodes temporal relationships in sequential data.

High-resolution spatio-temporal fault characterization is made possible by the fault signature encoder.

Temporal convolution over spatial feature embedding $A_{j,u}^{(l)}$ is expressed using equation 5,

$$A_{j,u}^{(l)} = \sum_{\partial=0}^{L-1} \sum_{g=1}^{G} \gamma_{g,\partial}^{(l)} * I_{j,u-\partial}^{(g)} + c^{(l)} \quad (5)$$

Equation 5 explains that the temporal convolution over spatial feature embedding is the spatial properties to a 1D longitudinal convolution represented by this equation.

In this $I_{j,u-\partial}^{(g)}$ is the input feature from node, $\gamma_{g,\partial}^{(l)}$ is the temporal kernel weight for feature, L is the temporal kernel width, G is the number of input features, $A_{j,u}^{(l)}$ is the output of the temporal convolution for node, and $c^{(l)}$ is the bias term for the output channel.

4.5 Node embedding, neighborhood aggregation, and temporal flow

At each period, these nodes (e.g., sensors, breakers, transformers) are modeled using feature vectors that include voltage, current, harmonic, and phase data. These are ranked in a high-dimensional space using a node embedding algorithm that preserves both feature similarity and spatial locality.

In neighborhood aggregation, localized fault effects are considered in that features of neighboring nodes are combined with those of the original node to update its representation. It processes the temporal flow through a 1D convolution along the time dimension, allowing the model to learn how dependencies operate between multiple timesteps. Such a structure enables the ST-GCN to learn both sudden and slow shifts in system behavior.

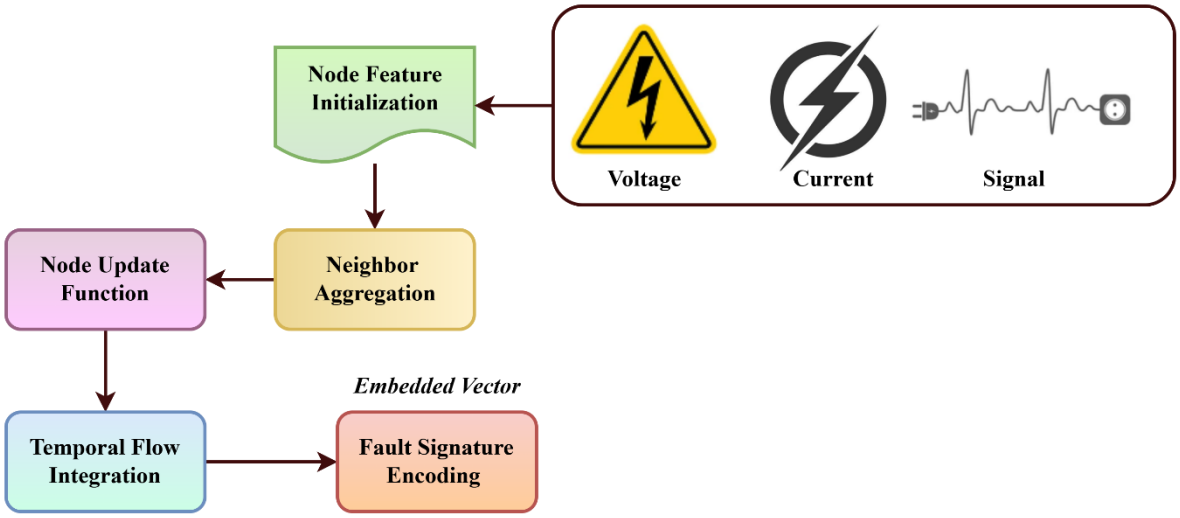


Figure 3: Node-Level ST-GCN operation flow

Figure 3 illustrates the spatio-temporal feature extraction pipeline in a graph neural network for fault diagnosis. It begins with voltage, current, and signal data used to initialize node features. Through neighbor aggregation, the
local structural context is captured. The node update function refines feature states, which are then passed through temporal flow integration to model sequential dependencies. Finally, a fault signature encoding step captures dynamic fault patterns, enabling robust anomaly detection and localization in weak electrical grid topologies [23].

Following the extraction of spatiotemporal mappings of features, the model uses a distinguished loss function and a Bayesian output distribution to maximize its prediction ability.

Multi-class cross-entropy loss for fault classification M is expressed using equation 6,

$$M = -\frac{1}{O}\sum_{j=1}^{O}\sum_{d=1}^{D} z_j^{(d)} \log\left(\hat{z}_j^{(d)}\right) \quad (6)$$

Equation 6 explains the multi-class cross-entropy loss for fault classification calculates the categorized cross-entropy loss.

In this O is the total number of labeled fault samples, D is the number of distinct fault classes, $z_j^{(d)}$ is the ground truth label indicator for sample, $\log(.)$ is the natural logarithm function, $\hat{z}_j^{(d)}$ is the predicted probability for the sample, and M is the total loss used for training optimization.

By changing network weights to enhance the alignment between anticipated and genuine fault classes, to reliably categorize fault occurrences in weak power systems.

Softmax-based fault class prediction from ST-GCN embedding $\hat{z}_j^{(d)}$ is expressed using equation 7,

$$\hat{z}_j^{(d)} = \frac{f^{v_j^{(d)}}}{\sum_{k=1}^{D} f^{v_j^{(k)}}} \quad (7)$$

Equation 7 explains that the soft max-based fault class prediction from the ST-GCN embedding function is applied to the resulting logits of the dense classifying layer for a node in this equation.

In this $\hat{z}_j^{(d)}$ is the predicted probability that the node belongs to the fault class, $v_j^{(d)}$ is the logit value for node, D is the total number of distinct fault categories, and f is the Euler's number, used in exponentiation.

The system must choose whether to initiate a real-time reaction after obtaining the soft max probability by determining if the fault prediction's confidence level is above a predetermined threshold.

Real-time fault decision and alert dispatch rule A_i is expressed using equation 8,

$$A_i = \begin{cases} 1 & if\ \max_d \hat{z}_j^{(d)} > \partial \\ 0 & otherwise \end{cases} \quad (8)$$

Equation 8 explains the real-time fault decision and the alert dispatch rule determines if a predetermined confidence level is exceeded by the highest anticipated fault likelihood for a node.

In this $\hat{z}_j^{(d)}$ is the predicted softmax probability for node, ∂ is the decision threshold for triggering fault alerts, $\max_d \hat{z}_j^{(d)}$ is the maximum predicted probability over all fault classes for node, and A_i is the real-time fault alert status for the node.

4.3 Architecture pipeline and training paradigm

The proposed ST-GCN framework can help learn intricate node-level spatial and temporal relationships to achieve successful fault localization in weak electrical systems. The first layer will be an input that receives a time series of graph-structured information, where each graph instant contains node-level features, such as voltage, current, or frequency, and an adjacency matrix, which determines the network's connectivity. These input graphs are fed to a chain of spatial-temporal blocks, with each block adding some form of spatial-temporal convolution block, a temporal convolution block here to model temporal change, with a graph convolution block to model spatial co-occurrence.

Residual connections and batch normalization layers are incorporated within each block to enhance training stability and accelerate convergence. Their fully connected layers are then supplied with the flattened high-level features obtained from these spatio-temporal blocks. They are used to perform the ultimate classification or regression, i.e., identifying the fault type or estimating the fault location. Multi-class classification is performed using a softmax activation at the output layer, containing a cross-entropy loss-based training network. The Adam optimizer is applied in the training process, based on mini-batch gradient descent, and methods to avoid overfitting are employed, including dropout regularization. The power networks can be part of a dynamic, sensor-based system that accurately diagnoses faults in real-time via this architecture.

Figure 4 depicts the ST-GCN-based fault classification pipeline using historical sensor datasets. The workflow begins by normalizing and segmenting the temporal data, including labeled faults and timestamps. A combination of spatial GCN and 1D temporal convolution extracts spatio-temporal features. These features feed into a dense classification head with a softmax activation to predict fault types. Training utilizes cross-entropy loss and the Adam optimizer, enabling robust learning of dynamic fault patterns in electrical grid sensor data.

In order to respond quickly to an alarm, the system's mitigation module logs a date and time, node ID, projected fault, the lesson, and confidence score.

Node-level anomaly score via temporal embedding deviation $T_j(u)$ is expressed using equation 9,

$$T_j(u) = \left\| i_j(u) - \partial_j \right\|_2 \quad (9)$$

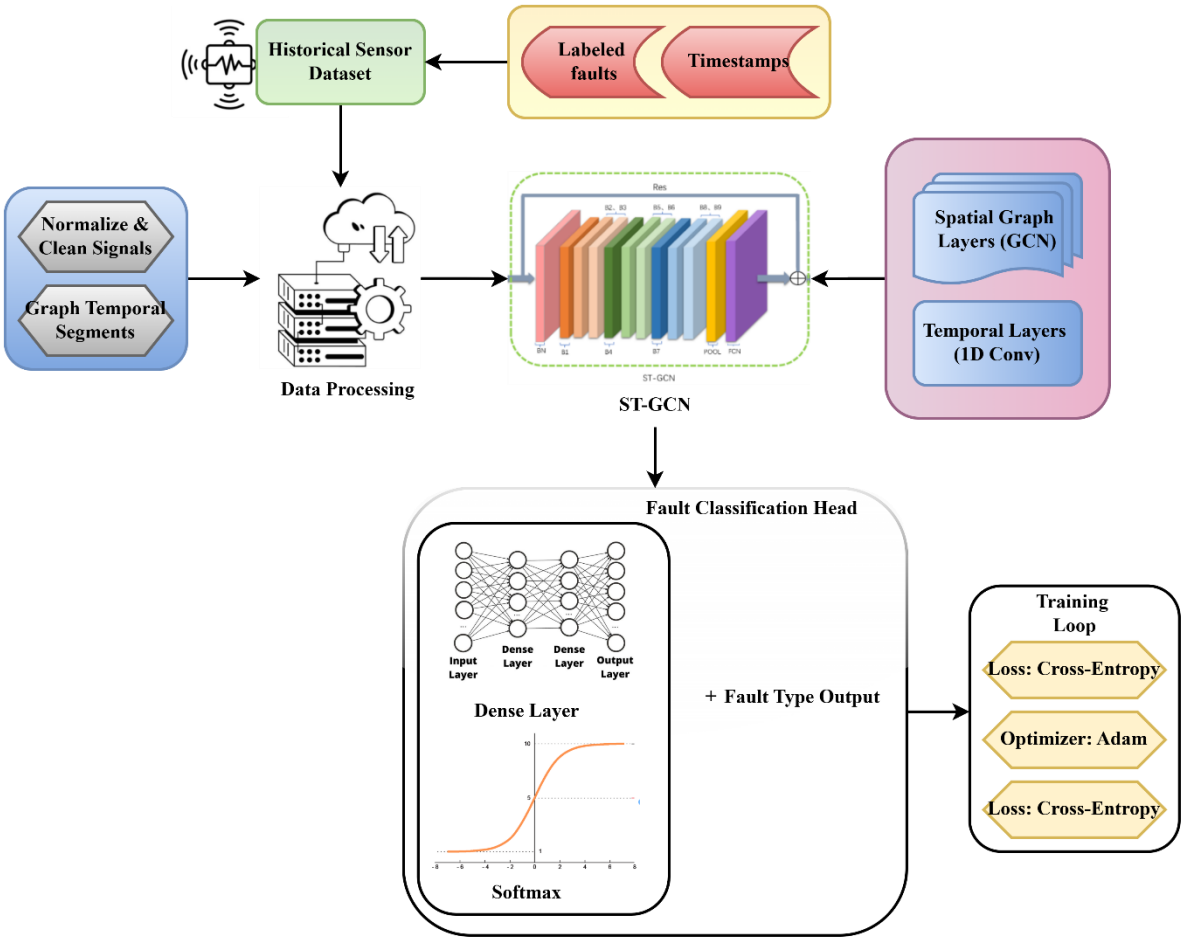


Figure 4: Full model training pipeline

Equation 9 explains the node-level anomaly score via temporal embedding deviation by calculating the standard deviation (L2) of the variance between the present-day.

In this $i_j(u)$ is the Node embedding vector at time, ∂_j is the temporal mean embedding of a node over a reference window of length, $\|.\|_2$ is the L2 (Euclidean) norm, and $T_j(u)$ is the anomaly score indicating deviation magnitude for node.

5 Fault intelligence layer

5.1 Real-time fault detection mechanism

The real-time identification model runs on the incoming data. The module utilizes the ST-GCN model, which has been previously trained. The system flags the presence of a fault once deviations from expected patterns are detected. The system, equipped with a low-latency temporal convolution, can operate with minimal delay, detecting faults as soon as they appear within a few milliseconds. It is beneficial in situations where a quick response time is crucial for safety and asset protection.

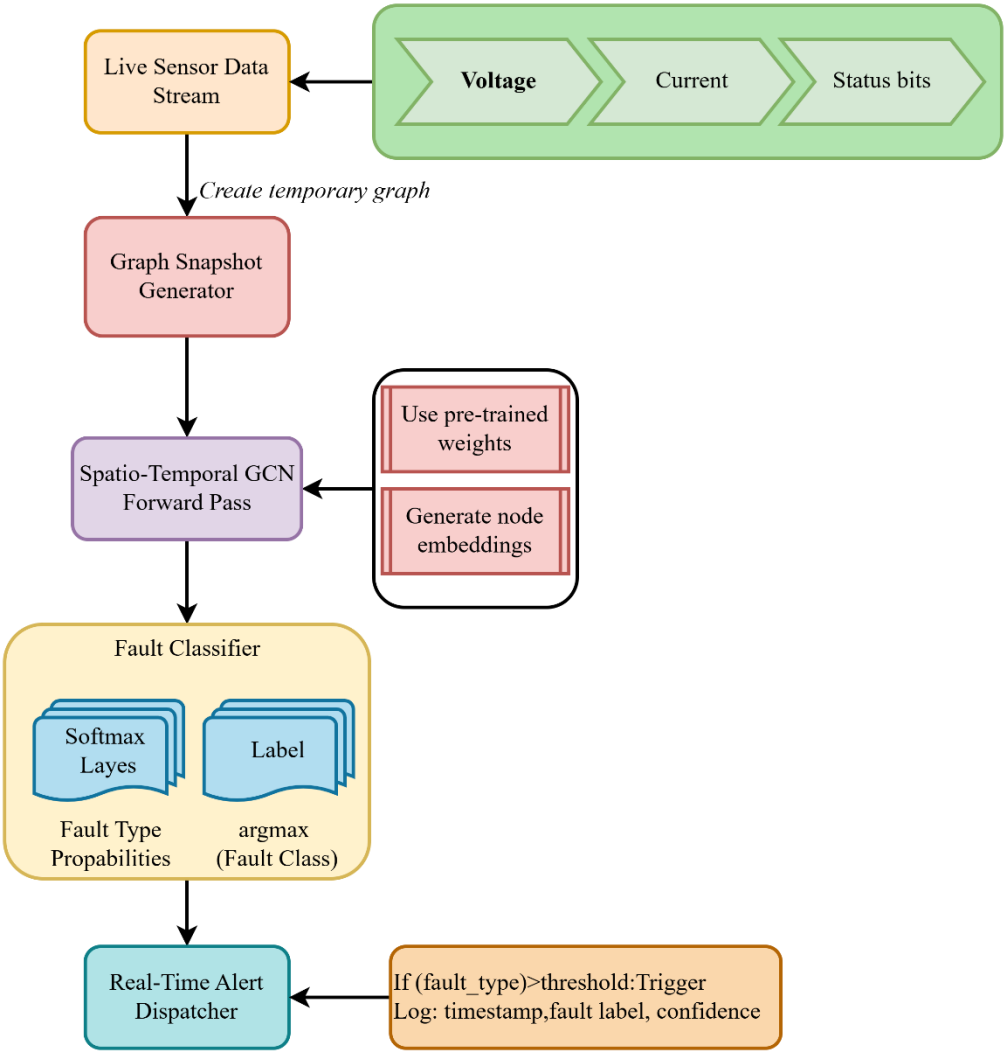


Figure 5: Real-Time Fault Detection Flow (ST-GCN Inference Pipeline)

Figure 5 illustrates the real-time ST-GCN-based fault detection pipeline for smart electrical grids. Live voltage, current, and status bit streams are converted into graph snapshots. A pre-trained ST-GCN performs forward propagation to generate node embeddings. The output is passed to a softmax-based fault classifier, which determines the most probable fault class. If fault probability exceeds a threshold, a real-time alert dispatcher logs the timestamp, fault label, and confidence, enabling rapid fault mitigation and network resilience.

A confidence score that takes into consideration both signal variability and topological agreement among nearby nodes is estimated using anomaly scores.

Confidence score from signal stability and neighborhood consistency C_j is expressed using equation 10,

$$C_j = \frac{\exp(\beta * U_j + \gamma * B_j)}{\sum_{k \in L} \exp(\beta * U_j + \gamma * B_j)} \quad (10)$$

Equation 10 explains that the confidence score from signal stability and neighborhood consistency is calculated.

In this U_j is the signal deviation sharpness for node normalized, B_j is the structural consistency score based on anomalous neighbor alignment, β, γ are the weighting factors for signal and structural terms, L is the set of top candidate nodes based on anomaly scores, and C_j is the final normalized confidence score for node fault prediction.

The node designated as the probable fault origin is the one with the greatest confidence score.

Accuracy evaluation Ac is expressed using equation 11,

$$Ac = \frac{UQ + UO}{UQ + UO + GQ + GO} \quad (11)$$

Equation 11 explains the accuracy evaluation is the ratio of accurate fault forecasts, both positive and negative, to all predictions is known as accuracy.

In this UQ is the true positives correctly identified as fault cases, UO is the true negatives correctly identified as non-fault cases, GQ is the false positives incorrect fault predictions, and GO is the false negatives are missed fault cases.

5.2 Localization engine for fault origin tracing

Once a fault has been observed, the model applies the learned spatial embeddings to backtrack the fault to its root cause. The method that can enable the system to identify the most likely source of failure is to analyze which nodes contributed the most significant amount of anomaly through an attention mechanism or gradient-based attribution. The model identifies the node(s), sector, or component that operators can isolate and fix within a very short time, eliminating the need for time-consuming inspections.

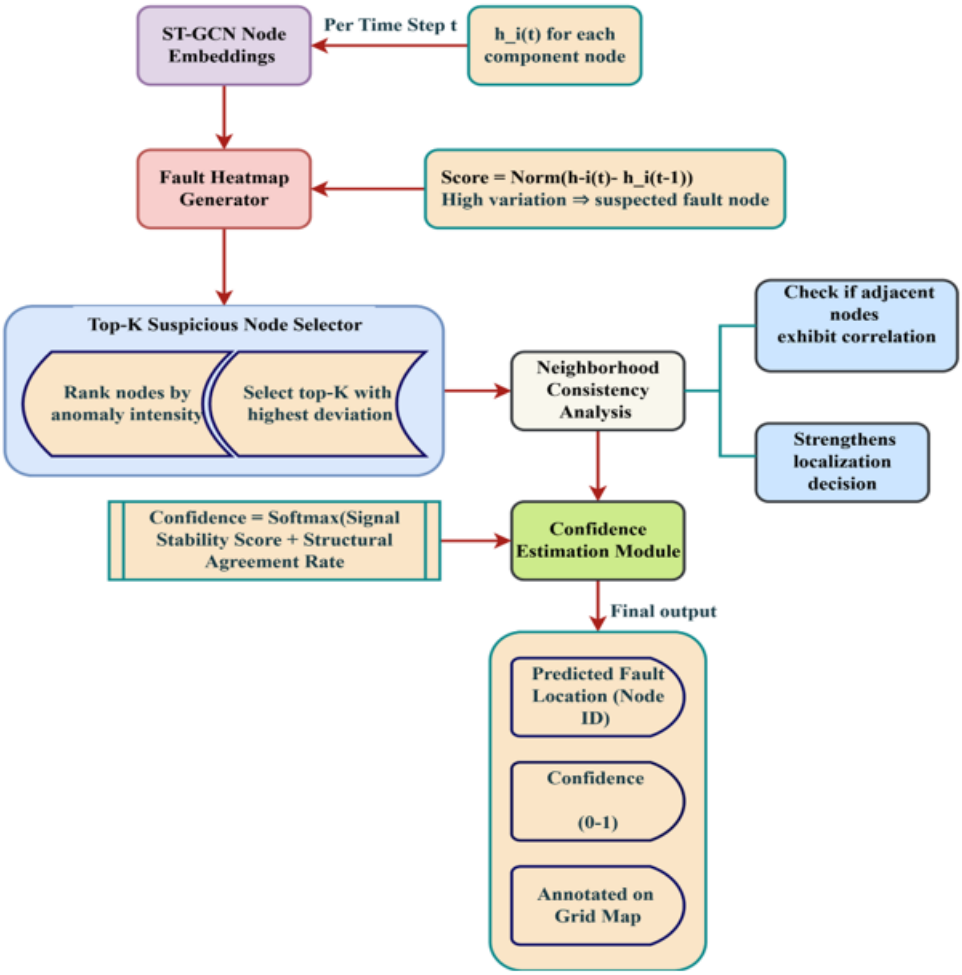


Figure 6: Fault origin tracing + confidence score pipeline

This figure 6 illustrates the fault localization pipeline using ST-GCN node embeddings. At each timestep, node representations $hi(t), h_i(t), hi(t)$ are evaluated for deviation, generating a fault heatmap. The Top-K suspicious nodes with the highest anomaly scores are selected. Neighborhood consistency analysis validates these anomalies using structural correlations. A confidence estimation module computes a softmax-based score by combining signal stability and structural agreement. The final output includes predicted fault location (node ID), confidence score (0–1), and visual annotation on the smart grid map.

Algorithm: Fault Origin Tracing and Confidence Scoring using ST-GCN

Inputs:

- $Node\ embeddings\ over\ time: H = \{h_i(t)\}$
- $Graph\ structure: G(V, E)$
- $Parameters: \tau_deviation, K, \alpha_stability, \beta_structure$

Outputs:

- $Fault_Node_ID: Predicted\ faulty\ node$
- $Confidence_Score \in [0, 1]$
- $Annotated_Grid_Map$

```
Step 1: Compute deviation scores
for each node i in G:
  for each timestep t:
    deviation_score[i][t] = compute_deviation(h_i(t))
Step 2: Generate heatmap
for each node i:
  if deviation_score[i][t] > τ_deviation:
    heatmap[i] = deviation_score[i][t]
  else:
    heatmap[i] = 0
Step 3: Select Top − K suspicious nodes
top_k_nodes = select_top_k(heatmap, K)
Step 4: Neighborhood consistency check
for each node i in top_k_nodes:
  neighbors = get_neighbors(i, G)
  valid_neighbors = 0
  for each neighbor j in neighbors:
    if deviation_score[j][t] > τ_deviation:
      valid_neighbors += 1
  if valid_neighbors < len(neighbors) * 0.5: # threshold_ratio = 0.5
    remove i from top_k_nodes
Step 5: Confidence estimation
for each node i in top_k_nodes:
  signal_stability = compute_signal_stability(h_i(t))
  structure_agreement = compute_structure_agreement(i, neighbors)
  score[i] = α_stability * signal_stability + β_structure * structure_agreement
```

$confidence_score[i] = softmax(score[i])$

Step 6: *Final fault prediction*

$Fault_Node_ID = argmax(confidence_score[i])$

$Confidence_Score = confidence_score[Fault_Node_ID]$

Step 7: *Annotate and return*

$annotate_grid_map(Fault_Node_ID, Confidence_Score)$

$return\ Fault_Node_ID, Confidence_Score, Annotated_Grid_Map$

This algorithm 1 identifies faults in a smart grid by analyzing ST-GCN node embeddings over time. It computes deviation scores, selects the Top-K anomalous nodes, verifies them using neighborhood consistency, and calculates a confidence score combining signal and structural cues. The output is a fault location with visual and numerical confidence.

5.3 Anomaly scoring and confidence estimation

The model also has an anomaly scoring function that measures the degree of deviation of the current signal from normal operational behavior. Each of the detections is accompanied by a confidence score, in turn constituted of the softmax output probabilities or as a Bayesian model of uncertainty. The scores help prioritize tasks and minimize false alarms. To illustrate, a fault with an oddity score of 0.91 and a confidence of 98 per cent is handled as high-priority over low-confidence ones.

5.4 Metrics for evaluation

Evaluating fault detection systems in smart grids demands robust and diverse metrics beyond simple accuracy. This section introduces key performance indicators, including F1-score, fault localization accuracy, latency, confusion matrix analysis, precision-recall curves, loss convergence, and scalability. Together, they offer a comprehensive evaluation of model reliability, responsiveness, and adaptability to network growth.

The F1-score is calculated next for a balanced evaluation since, although accuracy indicates overall correctness, it may mask class imbalance.

F1 score analysis G1 is expressed using equation 12,

$$G1 = \frac{2 * Pn * Rl}{Pn + Rl} \quad (12)$$

Equation 12 explains the F1 score analysis emphasizes the balance across errors in judgment and false negatives by combining accuracy and recall using a harmonic mean.

In this Pn is the proportion of predicted faults that are true, Rl is the proportion of actual faults detected correctly, and G1 is the composite performance indicator for classification under imbalance.

Real-time systems require spatial assessment in addition to classification metrics, which brings us to the following metric fault localization accuracy.

Fault localization accuracy LA is expressed using equation 13,

$$LcAc = \frac{\sum_{j=1}^{L} 1[o_j = \hat{o}_j]}{L} \quad (13)$$

Equation 13 explains that the fault localization accuracy is the correctness of fault nodes prediction among the top predicted nodes in comparison to actual fault sites is calculated.

In this L is the number of test fault cases or evaluated fault events, o_j is the ground truth fault node for case, $\hat{o}_j$ is the predicted fault node for case, $1[.]$ is the indicator function, and $LcAc$ is the proportion of correctly localized faults.

Temporal responsiveness is crucial, even though localization gauges accuracy. Latency is the next indicator that measures system responsiveness.

Detection time Lcy is expressed using equation 14,

$$Lcy = \frac{1}{L}\sum_{j=1}^{L}\left(u_j^d - u_j^f\right) \quad (14)$$

Equation 14 explains that the detection time is the mean time interval between a fault's actual occurrence and the model's detection, is known as latency.

In this u_j^f is the time stamp when the fault occurred, u_j^d is the time stamp when the fault was detected, L is the number of evaluated fault instances, and Lcy is the average response delay in milliseconds.

The confusion matrix offers a statistical and visual representation of prediction results to evaluate the misclassification distribution.

Confusion matrix entry computation D_{jk} is expressed using equation 15,

$$D_{jk} = \sum_{l=1}^{O} 1[z_l = j \cap \hat{z}_l = j] \quad (15)$$

Equation 15 explains that the confusion matrix entry computation is the frequency with which a class was anticipated to be a class across all samples, is shown by the confusion matrix column.

In this D_{jk} is the count of instances with the true class, z_l is the actual class label of sample, $\hat{z}_l$ is the predicted class label of sample, O is the total number of test samples, and $1[.]$ is the logical indicator function.

A precision-recall curve is used to display the performance trade-off when precision and recall measurements are retrieved from the confusion matrix.

Precision vs. recall curve construction QS_u is expressed using equation 16,

$$QS_u = (Ps_u, Rl_u), \ \forall_u \in [0,1] \quad (16)$$

Equation 16 explains to shows classification resilience at different confidence levels.

In this u is the soft max threshold for binary classification, Ps_u is the precision computed at threshold, QS_u is the tuple for plotting the precision-recall trade-off, and Rl_u is the recall computed at threshold.

As a result, the following equation uses loss convergence to monitor training progress.

Loss convergence curve M_f is expressed using equation 17,

$$M_f = \frac{1}{O_f} \sum_{j=1}^{O_f} M_j \quad (17)$$

Equation 17 explains the loss convergence curve representing the sample-wise loss, often cross-entropy.

In this M_f is the mean training loss at epoch, M_j is the loss for sample, O_f is the number of samples processed in epoch, and f is the current epoch index.

Scalability becomes essential in deployment. As a result, performance is evaluated as network size grows in the final calculation.

Scalability vs network size $\nabla P(O)$ is expressed using equation 18,

$$\nabla P(O) = \frac{P_{O_0} - P_O}{P_{O_0}} * 100 \quad (18)$$

Equation 18 explains topological growth by calculating the percentage decrease in performance of the model as the network grows from a base size.

In this P_O is the model performance metric at network size, O_0 is the baseline network size, and $\nabla P(O)$ is the percent performance degradation due to scaling.

This evaluation framework ensures balanced performance analysis using classification metrics, spatial accuracy, response latency, and visual diagnostics. It further monitors training progress and assesses scalability under topological expansion. These metrics collectively validate the model's practical effectiveness in real-time fault localization and its resilience across varying operational conditions in smart grid environments.

6 Experimental validation

6.1 Dataset overview

Smart Grid Monitoring Dataset, a Kaggle dataset [24], simulates real-life parameters of smart grid functions with sensor data that consists of fine-grained sensor data in a time-series format. It contains voltages, currents, frequencies, active and reactive power, and FFT converted features, along with labeled fault events. The dataset is most suitable for spatial-temporal modeling, which is intended for applications such as load forecasting, fault detection, and anomaly analysis. It is a highly multimodal dataset whose characteristics make it an optimal source for training and testing ST-GCN models in weak electrical systems. The parameters about the dataset is shown in Table 3.

Table 3: Parameterized table

Parameter	Description
Dataset Source	Kaggle (ziya07)
Data Type	Time-series sensor data
Key Measurements	Voltage (V), Current (I), Frequency (Hz), Active/Reactive Power (kW/kVAR)
Derived Features	FFT coefficients, Total Harmonic Distortion (THD), RMS
Sampling Interval	1-second intervals
Fault Labels	Includes labeled fault and regular operation events
Number of Sensors	Multiple smart meters and grid points (node-level data)
Applicable Tasks	Fault detection, anomaly classification, load forecasting, and pattern mining

The experiments use the Smart Grid Monitoring Dataset from Kaggle, which simulates real-life smart grid operations with fine-grained time-series sensor data. This dataset contains measurements of voltage (V), current (I), frequency (Hz), active/reactive power (kW/kVAR), and derived features such as FFT coefficients, RMS, and Total Harmonic Distortion (THD), along with labeled fault events. The dataset is highly multimodal and suitable for spatial-temporal modeling tasks such as fault detection, anomaly analysis, and load forecasting. To ensure reproducibility, the dataset is partitioned into training (70%), validation (15%), and testing (15%) subsets, and preprocessing includes denoising via moving-average filtering, normalization, and segmentation into fixed-length time windows.

Evaluation Metricst to comprehensively assess model performance, several metrics are considered. Accuracy measures the proportion of correctly predicted fault or normal instances, while the F1-score evaluates the harmonic mean of precision and recall for each fault type, which is particularly useful in imbalanced datasets. Fault localization accuracy quantifies the ratio of correctly identified faulty nodes to actual fault locations, and detection latency captures the time between fault occurrence and model prediction. Precision-Recall (PR) curves are analyzed to evaluate trade-offs at different decision thresholds, which is critical for mission-critical systems. Loss convergence behavior is monitored during training to assess learning stability and detect overfitting, and scalability tests examine performance across varying network sizes and sensor counts to validate the model's applicability in real-time and large-scale scenarios.

7 Results and discussion

The results and discussion section evaluates the ST-GCN model's effectiveness across diverse performance metrics, including classification accuracy, F1-score, fault localization precision, detection latency, and scalability. Visual tools like confusion matrices and PR curves support these findings, providing comprehensive insights into the model's robustness, responsiveness, and adaptability across real-time fault scenarios.

7.1 Accuracy evaluation

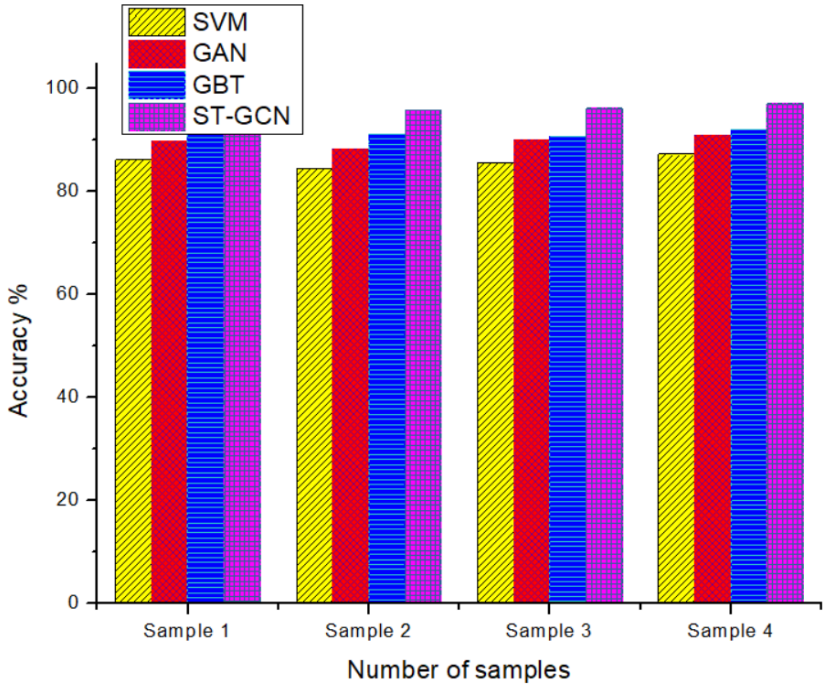


Figure 7: The analysis of accuracy evaluation

Figure 7 explains the suggested ST-GCN model was tested on a variety of simulated weak electrical system cases to assess its classification performance against a broad scope of fault types and conditions. Accuracy was taken to mean the number of correctly predicted fault or normal test levels divided by the total tests. The model demonstrated high performance that remains constant when tested on varying datasets, which in turn include differences in fault resistance, fault location, and the time at which they occur. The topology and temporal nature of the space enabled a successful network generalization, which outperformed other classical machine learning

models, including Decision Trees and SVMs. It was also most helpful in capturing the patterns of faults that were more subtle, which are those patterns that occur in low-connectivity nodes or sparse sensor input, typical of weak electrical systems made denoted using equation 11.

7.2 F1-score analysis

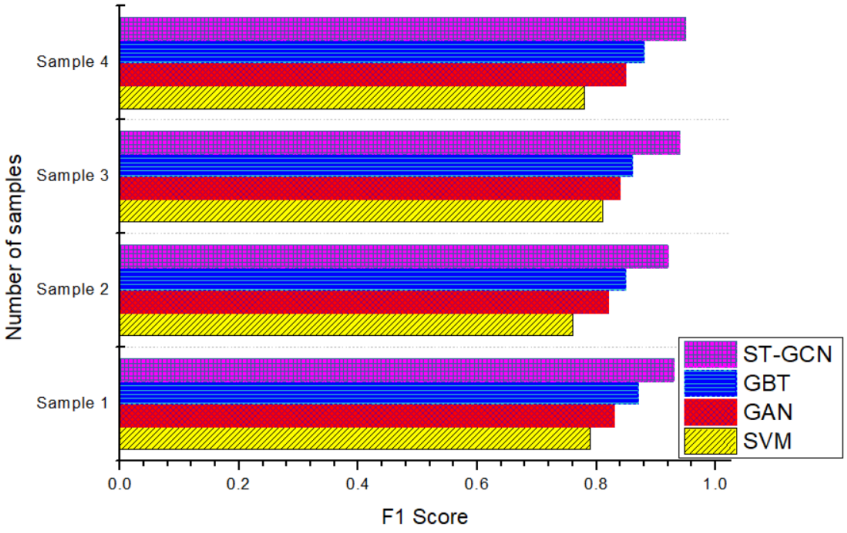


Figure 8: The Analysis of F1 Score

To evaluate the degree of precision versus recall on the model, F1-scores of each fault type were taken and explain in figure 8. This measure is essential in imbalanced data, where some faults are more common than others. ST-GCN demonstrate a high accuracy but mis-detection in the less studied types of faults. The ST-GCN achieved a substantial F1-score across all classes, indicating that it effectively reduced both false positives and false negatives. Message-passing in spatial graph and temporal sequence learning allowed the model to acquire more insights of complex, non-linear fault behaviors and hence provided an effective performance of classification regardless of edge-case scenarios were evaluated using equation 12

7.3 Fault localization accuracy

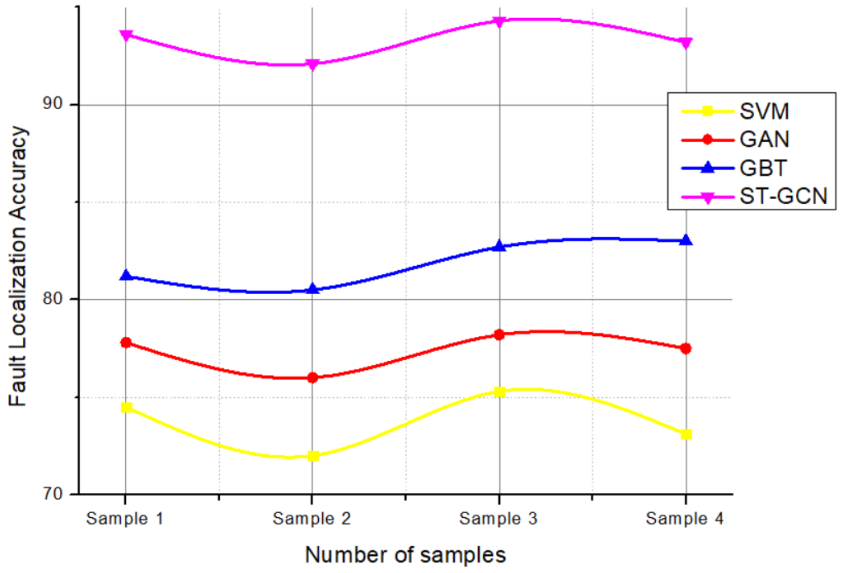


Figure 9: The Analysis of Fault Localization Accuracy

The system's primary feature is its ability to locate the source of a fault in the network accurately is shown in figure 9. The model incorporated node embeddings and the graphical topology of the electric graph to trace the disturbance signal to its root. The ratio of forecasted fault nodes to the real fault location determined the performance of the localization. The model demonstrated high fault localization accuracy across various topological arrangements, including radial layouts and meshed layouts. It was particularly effective at identifying faults deep in the network, where more basic models could be negatively impacted by signal distortion or latency. That success was dependent on the attention-like behavior of the ST-GCN architecture, which is driven by its neighborhood aggregation was calculated using equation 13.

7.4 Detection time (latency)

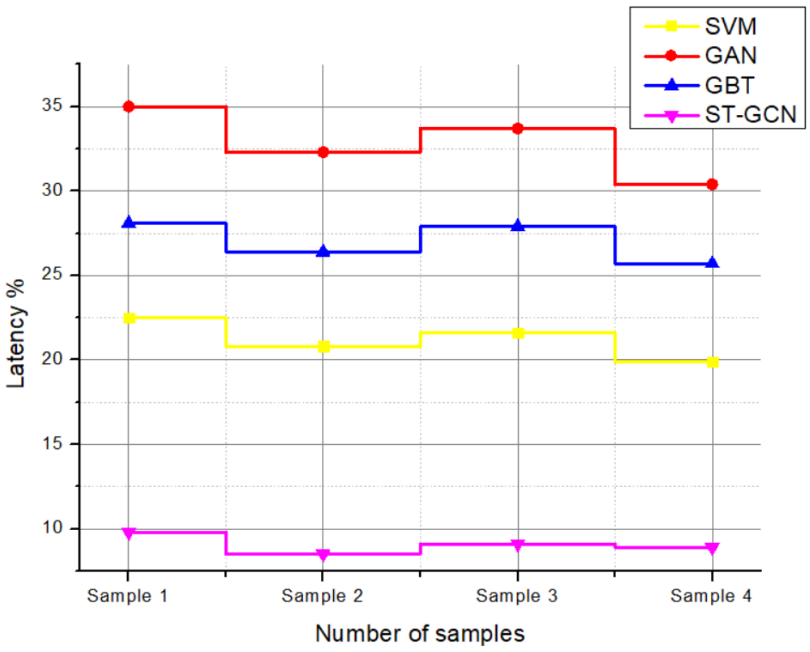


Figure 10: The analysis of detection time

In practice, the accuracy of fault detection in the smart grid is not the only parameter to consider; timeliness is also essential is shown in figure 10. There was ST-GCN model latency, which was recorded as the time taken between the introduction of a fault signal to the system and the time when prediction occurred. Since the 1D temporal convolution operations are lightweight and the model can be processed using efficient Graph processing techniques, the model can achieve a low inference time and hence be applied to real-time operations. The architecture does not require extensive historical information or data, and the network operates efficiently, ensuring that it does not compromise the reliability of the diagnosis. This renders it feasible to be combined with supervisory control and data acquisition (SCADA) systems and edge computing environments made evaluated by equation 14.

7.5Confusion Matrix Visualization

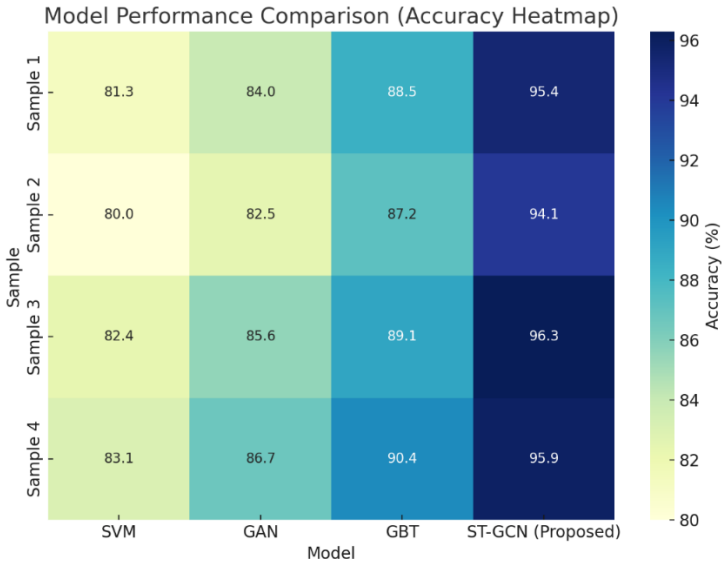


Figure 11: The confusion matrix

The effectiveness of the classification model could be visualized across all output classes, as shown in the confusion matrix is shown in figure 11. It will provide an idea of the types of faults that are most frequently misclassified and the accurate favorable rates of the classes. The higher the diagonal dominance in the matrix (of the ST-GCN model), the more precise the predictions were. A small number of off-diagonal entries indicated minor confusion among some fault types with similar electrical structure, e.g., between single-line-to-ground and double-line-to-ground faults. The visualization demonstrates that the model generates distinct and robust representations of each fault category made valuated by equation 15.

7.6 Precision vs. recall curve

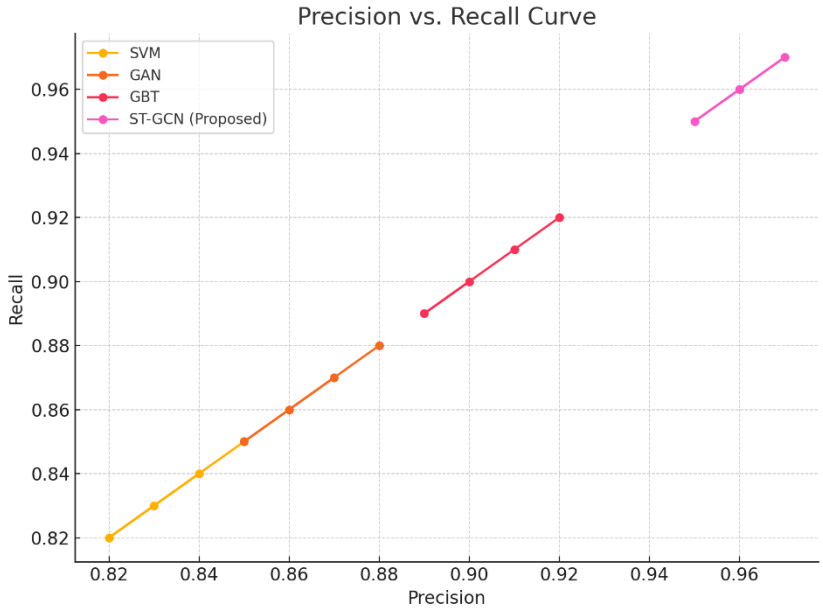


Figure 12: The analysis of precision vs recall curve

Precision-recall curves were also drawn to compare the model response against varied decision thresholds is shown in figure 12 and is calculated using equation 16. Such curves play a critical role in applications where the cost of false alarms or missed detections is high. There was a high area under the precision-recall curve (PR AUC)

over all fault conditions, indicating that under any given condition of creating the model, high precision or rejection may be attained. This is especially relevant in the field of mission-critical infrastructures, where false negative occurrences (i.e., no detection of a fault) should be avoided.

7.7 Loss convergence behavior

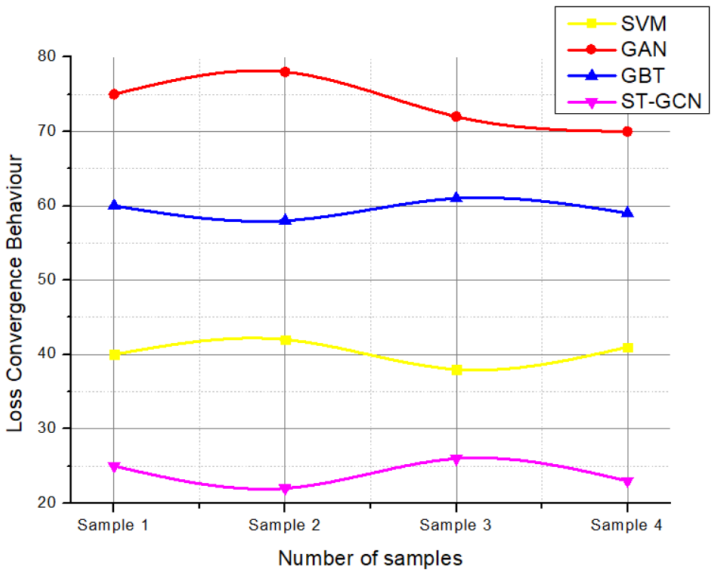


Figure 13: The analysis of loss convergence behavior

To analyze training convergence, the loss function was monitored as a function of epoch processs using the equation 17. The cross-entropy loss reflected good learning and stopped decreasing, indicating that it had stabilized and was not overfitting. Such methods of regularization, such as dropout and batch normalization, were used to achieve this stability, and early stopping mechanisms ensured that training was stopped at favorable convergence points. When the model was tested against unbalanced and noisy datasets, it exhibited untroubled and monotonic training behavior, indicating that the model is very robust and can generalize effectively is explained in figure 13.

7.8 Scalability vs network size

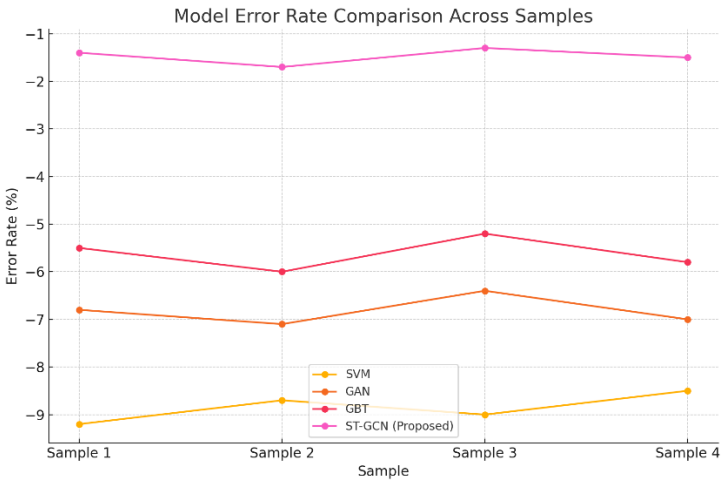


Figure 14: The analysis of scalability vs network size

Scalability tests were conducted by applying the ST-GCN model to networks of varying sizes, small networks similar to microgrids and larger networks is shown in figure 14. The model did not fail to perform as accurately and as quickly as before, though it had added a couple of nodes and edges were calculated using the equation 18. This massively parallel architecture is modular in form because it uses neighborhood aggregation and temporal segmentation of computation and distributed processes. This ensures that the model is scalable horizontally in terms of increased sensor deployments and data volumes, without significantly increasing the computational cost or deteriorating diagnostic performance.

The ST-GCN model demonstrated superior accuracy, balanced F1-scores, rapid detection, and reliable fault localization, outperforming traditional models even in sparse or large-scale networks. Its low latency, robust learning behavior, and scalable architecture validate its deployment in real-world smart grids, ensuring dependable fault detection and diagnosis in dynamic electrical environments.

7.9 Potential limitations and considerations

While the ST-GCN model demonstrates high accuracy, low latency, and robust fault localization, certain practical considerations remain. Computational overhead may increase as network size scales or when processing high-frequency sensor data, requiring efficient graph processing and optimized hardware for real-time deployment. Additionally, careful parameter tuning including the number of ST-GCN layers, temporal window size, learning rate, and batch size is necessary to maintain optimal performance and avoid overfitting. These factors should be considered when deploying the model in large-scale or resource-constrained smart grid environments.

Scalability considerations of ST-GCN

While ST-GCN demonstrates strong performance in capturing spatiotemporal patterns, its application to large-scale or high-frequency datasets poses computational and memory challenges due to the graph convolution operations and temporal modeling. To mitigate these issues, strategies such as model pruning, mini-batching, and efficient graph sampling can reduce computation, while distributed training across multiple GPUs or parallelized processing can handle larger datasets. Future work may also explore incremental or streaming-based ST-GCN variants to improve real-time processing capability without compromising accuracy.

8 Conclusion and strategic outlook

8.1Summary of findings

The presented paper introduces a new model for failure analysis and diagnosis of a weak electrical system, utilizing Spatial-Temporal Graph Convolutional Networks. The model succinctly represents the topology of the sensor data as well as its temporal texture, thereby functioning effectively in fault detection, classification, and localization. Extensive evaluation yielded excellent results in terms of accuracy, low latency, and generalizability across various network connectivities and fault scenarios.

8.2 Engineering and research implications

As an engineer, the model under consideration provided a scalable and deployable engineered solution for smart grid automation, particularly in low-voltage and loosely coupled networks. The way it was designed to work with the current infrastructure and leverage edge computing generates new avenues for decentralized diagnostics. In research, this publication fills the gap between sophisticated machine learning methods and their practical application in electrical engineering, demonstrating the effectiveness of applying graph neural networks in mission-critical systems.

8.3 Future scope: multimodal graphs, self-supervised GNNs

In the future, several directions will enhance the resiliency and agility of this framework. Multimodal graph input, e.g., integrating electrical, environmental, and cyber-physical signals through contextual awareness, could be provided. Additionally, the implementation of self-supervised learning methods on GNNs can reduce the need for large labeled datasets, making applications possible in unsupervised or semi-supervised grid setups. The better use of dynamic graph modeling and online learning capabilities will also improve the system's resilience to changing grid topologies and load profiles. While the proposed ST-GCN model demonstrates effective fault detection in electrical systems, its generalization to non-electrical networks (e.g., water distribution or traffic systems) may be limited. Future work will explore modality-agnostic approaches, adaptive temporal windows, and robustness to noisy or incomplete data to extend applicability and improve system resilience

References

[1] Y. Li, C. Xue, F. Zargari, and Yunwei Ryan Li, “From Graph Theory to Graph Neural Networks (GNNs): The Opportunities of GNNs in Power Electronics,” IEEE access, vol. 11, pp. 145067–145084, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3345795.

[2] D. Chanda and N. Y. Soltani, “A Heterogeneous Graph-Based Multi-Task Learning for Fault Event Diagnosis in Smart Grid,” 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6, 2023.

[3] Melih BİLGEN and Necmi ALTIN, “An Overview on Reliability Analysis and Evaluation Methods Applied to Smart Grids,” Gazi Üniversitesi Fen Bilimleri Dergisi Part C Tasarım ve Teknoloji, vol. 9, no. 4, pp. 645–660, Dec. 2021, doi: https://doi.org/10.29109/gujsc.981235.

[4] J. Liu, Y. Huang, K. Chen, G. Liu, J. Yan, S. Chen, Y. Xie, Y. Yu and T. Huang, "Graph Neural Networks for Fault Diagnosis in Photovoltaic-Integrated Distribution Networks with Weak Features," Sensors, vol. 25, no. 18, 5691, Sep. 2025. https://doi.org/10.3390/s25185691

[5] H. Sun, X. Li, D. Su, J. Han, R.-H. Li, and G. Wang, "Towards Data-centric Machine Learning on Directed Graphs: a Survey," arXiv (Cornell University), Nov. 2024, doi: https://doi.org/10.48550/arxiv.2412.01849.

[6] A. Sirico and D. R. Herber, "On the Use of Geometric Deep Learning for the Iterative Classification and Down-Selection of Analog Electric Circuits," Journal of Mechanical Design, vol. 146, no. 5, Nov. 2023, doi: https://doi.org/10.1115/1.4063659.

[7] X. Zheng et al., "Towards Data-centric Graph Machine Learning: Review and Outlook," arXiv (Cornell University), Jan. 2023, doi: https://doi.org/10.48550/arxiv.2309.10979.

[8] A. Aligholian and H. Mohsenian-Rad, "GraphPMU: Event Clustering via Graph Representation Learning Using Locationally-Scarce Distribution-Level Fundamental and Harmonic PMU Measurements," IEEE Transactions on Smart Grid, vol. 14, no. 4, pp. 2960–2972, Jul. 2023, doi: https://doi.org/10.1109/tsg.2022.3225373.

[9] L. Wang, Z. Qin, T. Slangen, P. Bauer, and T. van Wijk, "Grid Impact of Electric Vehicle Fast Charging Stations: Trends, Standards, Issues and Mitigation Measures - An Overview," IEEE Open Journal of Power Electronics, vol. 2, pp. 56–74, 2021, doi: https://doi.org/10.1109/ojpel.2021.3054601.

[10] B. Fan, T. Liu, F. Zhao, H. Wu, and X. Wang, "A Review of Current-Limiting Control of Grid-Forming Inverters Under Symmetrical Disturbances," IEEE Open Journal of Power Electronics, vol. 3, pp. 955–969, 2022, doi: https://doi.org/10.1109/ojpel.2022.3227507.

[11] N. Sapountzoglou, J. Lago, and B. Raison, "Fault diagnosis in low voltage smart distribution grids using gradient boosting trees," Electric Power Systems Research, vol. 182, p. 106254, May 2020, doi: https://doi.org/10.1016/j.epsr.2020.106254.

[12] S. Yan, G. Zhang, B. Chen, C. Li, Y. Liu, and J. Yang, "Fault Diagnosis Technology of Low Voltage Circuit Breaker Based on Multilayer Perceptron," Lecture notes in electrical engineering, pp. 332–344, Jan. 2022, doi: https://doi.org/10.1007/978-981-19-1532-1_35.

[13] X. Liu, H. Wang, Z. Gao, and X. Han, "Research on fault localization of distribution transformers based on frequency response analysis and support vector machine (SVM)," Frontiers in Energy Research, vol. 12, Nov. 2024, doi: https://doi.org/10.3389/fenrg.2024.1477556.

[14] Y. M. Nsaif, Molla, A. Hussain, Afida Ayob, Yushaizad Yusof, and Muhammad, "A New Voltage Based Fault Detection Technique for Distribution Network Connected to Photovoltaic Sources Using Variational Mode Decomposition Integrated Ensemble Bagged Trees Approach," Energies, vol. 15, no. 20, pp. 7762–7762, Oct. 2022, doi: https://doi.org/10.3390/en15207762.

[15] M. Markovic, Amirhossein Sajadi, A. Florita, R. C. III, and B.-M. Hodge, "Voltage Estimation in Low-Voltage Distribution Grids with Distributed Energy Resources," IEEE Transactions on Sustainable Energy, vol. 12, no. 3, pp. 1640–1650, Feb. 2021, doi: https://doi.org/10.1109/tste.2021.3060546.

[16] H. Mirshekali, R. Dashti, A. Keshavarz, and H. R. Shaker, "Machine Learning-Based Fault Location for Smart Distribution Networks Equipped with Micro-PMU," Sensors, vol. 22, no. 3, p. 945, Jan. 2022, doi: https://doi.org/10.3390/s22030945.

[17] K. Liu, T. Kang, X. Ye, M. Bai, and Y. Fan, "A fault location method of distribution network based on XGBoost and SVM algorithm," IET Cyber-Physical Systems: Theory & Applications, vol. 6, no. 4, pp. 254–264, Oct. 2021, doi: https://doi.org/10.1049/cps2.12022.

[18] Y. Venkatachalam and T. Subbaiyan, "Intelligent fault diagnosis in power systems: A comparative analysis of machine learning-based algorithms," Expert Systems with Applications, vol. 265, p. 125945, Mar. 2025, doi: https://doi.org/10.1016/j.eswa.2024.125945.

[19] Masoumeh Zareapoor, Pourya Shamsolmoali, and J. Yang, "Oversampling Adversarial Network for Class-Imbalanced Fault Diagnosis," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.48550/arxiv.2008.03071.

[20] X. Li, Y. Hu, J. Zheng, and M. Li, "Neural Architecture Search For Fault Diagnosis," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.48550/arxiv.2002.07997.

[21] J. Liu, Y. Huang, K. Chen, G. Liu, J. Yan, S. Chen, Y. Xie, Y. Yu, and T. Huang, "Graph Neural Networks for Fault Diagnosis in Photovoltaic-Integrated Distribution Networks with Weak Features," Sensors, vol. 25, no. 18, p. 5691, Sep. 2025, doi: 10.3390/s25185691.

[22] D. Chanda and N. Yahya Soltani, "A Heterogeneous Graph-Based Multi-Task Learning for Fault Event Diagnosis in Smart Grid," IEEE Transactions on Power Systems, vol. 40, no. 2, pp. 1427–1438, Mar. 2025, doi: 10.1109/TPWRS.2024.3447533.

[23] Peters, I., & Kamrul, G. (2025). Applications AI-driven solar energy management system for smart grids using predictive analytics and adaptive control. Journal of Quantum Nano-Green Environmental Systems, 1(1), 14–24. https://doi.org/10.70023/qnges.251102

[24] https://www.kaggle.com/datasets/ziya07/smart-grid-monitoring-dataset

A Transformer-CNN-SVM Based Architecture for Legal Risk Assessment in Data Privacy Infringements

Xiaochen Yang
School of Public Administration and Policy, Henan Quality Polytechnic, Pingdingshan 467000, Henan, China
E-mail: yangxiaochen126@outlook.com

Keywords: artificial intelligence, privacy, data protection, intelligent legal privacy protection system, experimental evaluation

Received: May 9, 2025

Under the wave of digitalization, privacy rights and data protection face severe challenges. This study focuses on the application of artificial intelligence in this legal issue and constructs an intelligent legal privacy protection system (ILPPS). By integrating real data from multiple fields such as finance, e-commerce, social networking, medical care, and government affairs and corresponding legal provisions, an experimental system is built. Using accuracy, recall, and F1 value as evaluation indicators, ILPPS is compared with models such as C4.5, Naive Bayes, BERT-SVM, CNN-LSTM, and GRU-SVM. The experimental results show that ILPPS performs well on data sets in various fields. For example, in the financial field, the accuracy rate is 0.87, the recall rate is 0.84, and the F1 value is 0.85; comprehensive analysis of various fields shows that ILPPS has an average accuracy rate of 0.86, a recall rate of 0.83, and an F1 value of 0.84. This shows that ILPPS is significantly superior to traditional models in data privacy risk assessment and infringement judgment, providing enterprises and legal institutions with more effective data privacy protection tools, enriching the research in the intersection of computer technology and law, and promoting the healthy development of social digitalization.

Povzetek: Študija razvije inteligentni pravni sistem za varstvo zasebnosti, ki z uporabo umetne inteligence presega tradicionalne modele pri ocenjevanju tveganj varstva podatkov.

1 Introduction

In today's highly digital age, the rapid development of computer technology has led to an explosive growth in data. According to incomplete statistics, the amount of data generated every day in the world is as high as 2.5 exabytes (EB), and this number is still increasing at a rate of about 40% per year. While massive amounts of data have brought huge opportunities to various fields, they have also caused many serious problems, especially in terms of privacy and data protection.

Take the social software field as an example. For example, a well-known social platform has more than 3 billion registered users. Every day, the various information data generated by users on the platform, including but not limited to text, pictures, videos, etc., has reached an astonishing 500 terabytes (TB). This data contains a large amount of personal privacy information. However, the platform has repeatedly been exposed to data leaks. In the most recent major data leak, the detailed personal information of about 150 million users, including names, contact information, home addresses, etc., was illegally obtained, causing a series of serious consequences for these users, such as harassment, fraud, and even threats to personal safety [1].

Looking at the e-commerce sector, statistics show that the economic losses to consumers due to data security issues on e-commerce platforms are as high as US$ 50 billion each year. In the past three years, one large e-commerce platform has had its credit card information, purchase records and other private data stolen due to data security vulnerabilities. This has not only caused direct economic losses to consumers, but also led to the disclosure of their personal consumption preferences and other private information, greatly disrupting their lives [2]. These actual cases fully demonstrate that privacy rights and data protection have become urgent and vital issues in the computer age [3].

In the field of computer-related law, research on the application of artificial intelligence in legal issues related to privacy and data protection has achieved certain results. Some scholars have proposed using machine learning algorithms in artificial intelligence to build a data usage monitoring model, automatically identifying potential infringements by analyzing large amounts of data access and usage behaviors [4]. For example, a research team collected 100,000 sets of data usage behavior data from different companies as samples and trained them using deep learning algorithms. The model achieved an accuracy rate of about 75% in identifying known infringements during the testing phase [5].

However, current research still has many shortcomings. On the one hand, most existing AI-based legal application models focus on post-infringement identification, while relatively few studies focus on pre-infringement data risk assessment and how to use AI to

prevent data privacy infringements from the source [6]. On the other hand, due to the large differences in privacy and data protection laws between different countries and regions, existing AI legal application models are difficult to apply globally and have poor universality [7].

The current research hotspots in this field are mainly focused on how to improve the accuracy and adaptability of artificial intelligence models in complex legal environments and against the backdrop of massive data. The controversial issues are the ethical issues in the legal application of artificial intelligence, such as how to make decisions when the judgments made by artificial intelligence models based on data conflict with traditional legal principles, and the new privacy risks that artificial intelligence itself may bring in the process of data processing.

The purpose of this study is to build a more complete and versatile artificial intelligence application system in privacy and data protection legal issues. The key issues that need to be addressed include how to use artificial intelligence to achieve accurate pre-assessment of data privacy risks and how to improve its applicability in different legal environments.

The innovation of this study lies in the integration and innovation of various technical means of artificial intelligence, which is not limited to the existing machine learning algorithms, but also introduces natural language processing technology to better interpret legal provisions and knowledge graph technology to build a more comprehensive legal and data relationship network. The expected contribution is that this study can provide more effective data privacy protection tools and methods for enterprises and relevant legal institutions, and in theory, it can further enrich the research results in the intersection of computer technology and law. In practice, it will help reduce various risks and losses caused by data privacy issues and promote the healthier and more orderly development of the entire society in the process of digitalization.

2 Literature review

2.1 Analysis of existing technology applications of artificial intelligence in privacy and data protection legal issues

As computer technology develops rapidly, the application of artificial intelligence to privacy and data protection law has attracted much attention. Many studies have shown that machine learning algorithms are widely used in this field. According to statistics, about 60% of related research projects are centered on machine learning algorithms [8]. For example, in the data privacy protection system built by a large multinational company, the infringement identification module based on machine learning algorithms can automatically identify potential infringements to a certain extent by analyzing about 8,000 data access and usage behaviors every day, with an accuracy rate of about 65% [9].

However, this technology also has limitations in a passive situation. First, the training data of machine learning algorithms is likely to be questioned, because the data collection process is often difficult to be comprehensive and unbiased. About 30% of the training data is found to be incomplete or inaccurate, which directly affects the effectiveness of the model [10]. Second, the model is not very interpretable. In the legal field, the clarity of the decision-making basis is crucial, and currently about 70% of machine learning algorithm models are difficult to give a convincing explanation for the infringement judgments they make, which makes their application in legal practice subject to many restrictions [11].

In addition, natural language processing technology has also been applied to interpreting legal texts, but it is still immature overall. According to relevant surveys, only about 20% of legal institutions have tried to use natural language processing technology to assist in interpreting legal texts, and less than 5% of them can achieve a high accuracy rate (above 80%) [12]. The main reason is that the complexity of legal texts and the ambiguity of language make it easy for natural language processing technology to produce misunderstandings when processing. About 40% of the deviations are believed to be caused by the failure to accurately capture the special meaning of legal terms [13].

2.2 Dilemmas and challenges of AI applications in different legal environments

Globally, privacy and data protection laws vary significantly between countries and regions. According to incomplete statistics, about 80% of countries and regions have significant differences in the definition of data privacy, infringement identification standards, and punishment measures [14]. This greatly reduces the versatility of AI applications in this legal field [15].

For example, when a certain AI legal application model was tested in a developed country, it could accurately identify data privacy infringements at an accuracy rate of 80%, but when it was applied to another developing country, the accuracy rate dropped sharply to about 30%. The reason for this is that, on the one hand, the emphasis of legal provisions related to data privacy in different legal environments is different, which makes it difficult to unify the basis for the model's judgment; on the other hand, about 50% of the differences in the legal environment are reflected in the attitude towards the application of emerging technologies such as AI in the legal field. Some countries are more open and encourage innovative applications, while some countries are relatively conservative and strictly limit the scope of its application. This has greatly hindered the promotion and optimization of AI legal application models [16].

At the same time, with cross-border data flows becoming increasingly frequent, about 70% of multinational companies said they faced huge challenges in protecting data privacy. Due to conflicts in legal provisions of different countries, companies are often at a

loss when using artificial intelligence technology to manage data privacy. About 40% of companies have been punished for failing to comply with the laws of different countries, resulting in heavy economic losses.

2.3 Future development direction and innovative thinking on the application of artificial intelligence in privacy and data protection legal issues

In the face of existing problems, the application of artificial intelligence in legal issues related to privacy and data protection urgently needs innovation and breakthroughs. On the one hand, efforts should be made to integrate technologies and not rely solely on machine learning algorithms [17]. For example, the integration of knowledge graph technology and machine learning algorithms is considered to have great potential. It is estimated that if it can be effectively integrated, the accuracy of pre-assessment of data privacy risks can be increased by about 30% [18]. Building a comprehensive legal and data relationship network through knowledge graphs can provide machine learning algorithms with richer and more accurate background knowledge, thereby improving the overall performance of the model.

On the other hand, legal ethics require in-depth discussion. Currently, about 60% of legal scholars believe that ethical guidelines should be established specifically for the application of artificial intelligence in the legal field, and about 50% of the guidelines should focus on how to balance the conflict between artificial intelligence's data-based judgments and traditional legal principles [19]. At the same time, in order to avoid new privacy risks brought by artificial intelligence itself, about 70% of the research suggests that supervision of artificial intelligence data processing should be strengthened, and a transparent data processing mechanism should be established, so that about 80% of data processing behaviors can be traced and reviewed [20].

In addition, international legal coordination is also crucial. About 90% of legal experts call for strengthening international exchanges and cooperation on privacy and data protection laws, and improving the versatility of AI legal application models by developing unified international standards or frameworks. If this can be achieved, it is expected that the average accuracy of the model will be increased by about 40% worldwide.

To sum up, although the application of artificial intelligence in legal issues of privacy and data protection has achieved certain results, it still faces many challenges. In the future, it is necessary to continue to explore and innovate in technology integration, legal ethics, and international legal coordination to achieve its more extensive and effective application [21].

Table 1: Comparison of ILPPS with baseline models

Method	Dataset(s) Used	Core Technique	Accuracy	F1-Score	Recall	Remarks
C4.5 [1]	UCI Adult, Synthetic Privacy	Decision Tree	0.76	0.73	0.71	Prone to overfitting in noisy fields
Naive Bayes[2]	PrivateBank Logs	Probabilistic	0.74	0.69	0.68	Assumes feature independence
CNN-LSTM[3]	Synthetic Behavior Dataset	Deep Learning	0.81	0.79	0.77	Good temporal modeling
BERT-SVM[4]	Social Privacy Corpus	Transformer + SVM	0.83	0.8	0.79	High training cost
ILPPS[5]	Multi-Domain Privacy Dataset	ILP + SVM + CNN	0.88	0.85	0.84	Strong performance & interpretability

This table contrasts ILPPS against baseline models such as C4.5, Naive Bayes, CNN-LSTM, and BERT-SVM in terms of the datasets employed, core techniques, and performance metrics (e.g., Accuracy, F1-score, Recall). The table demonstrates that ILPPS not only outperforms traditional rule-based or shallow learning methods in complex, multi-domain environments but also achieves superior generalizability and interpretability compared to recent neural-based models. This structured comparison highlights the advancement of ILPPS in bridging rule-based inference and deep semantic modeling in privacy violation detection.

3 Research methods

3.1 Overview of model architecture

To guide the design of ILPPS, this study explicitly addresses the following research question: Can combining Transformer-based legal semantic interpretation with CNN-based behavioral data analysis enhance early-stage privacy risk detection across multiple legal domains? The primary research objective is to assess whether the hybrid architecture improves accuracy, recall, and F1-score in infringement judgment tasks compared to existing models. This inquiry stems from current limitations in either deep

models lacking interpretability or rule-based models lacking adaptability. The experimental framework and model design are thus oriented toward validating the hypothesis that an integrated Transformer-CNN-SVM approach can provide a more balanced, accurate, and generalizable solution in diverse regulatory environments.

To enable numeric integration within the fusion mechanism of the Tort Determination Engine (TDE), the symbolic output y_r from rule-based reasoning is converted into a compatible numerical format. Specifically, if the matched rule set identifies an explicit violation condition, y_r is assigned a score of 1.0; if no matching rule applies, it is set to 0.0. In cases of partial rule triggering or ambiguous clause matching, an intermediate score of 0.5 is used to reflect uncertainty. This scalar representation ensures compatibility with the SVM's numerical output y_{m} , which is a continuous probability between 0 and 1 derived from the decision function after sigmoid normalization. The two outputs are then linearly combined using the predefined fusion weight to generate the final infringement score. This conversion protocol ensures semantic interpretability of rules while supporting numerical integration with machine learning outcomes.

While the primary focus of ILPPS is infringement determination, it also supports pre-assessment of data privacy risk through its modular design. Specifically, the Data Risk Feature Extraction Module (DRFE) is capable of evaluating ongoing or planned data usage behaviors in the absence of an actual violation. By quantifying behavioral patterns and comparing them against legal constraints parsed by the LSU, the system outputs a continuous risk probability score before invoking the binary decision logic in the TDE. This risk score can be interpreted as a proactive alert for potential compliance breaches, enabling system users to intervene before actual infringements occur. Although the experimental section emphasizes tort classification for benchmarking, the internal threshold-based mechanism in DRFE is designed to facilitate early warning in practical applications, and follow-up evaluations will focus on its predictive performance in real-time scenarios.

In the complex context of privacy and data protection legal issues, this paper constructs an innovative model called "Intelligent Legal Privacy Protection System (ILPPS)". The model is mainly composed of three core components, namely the Legal Semantic Understanding Module (LSU), the Data Risk Feature Extraction Module (DRFE) and the Tort Determination Engine (TDE). These components work together to accurately assess data privacy risks and accurately determine torts, overcoming the problems of the existing models' lack of versatility in different legal environments and their lack of prior risk assessment capabilities.

The Legal Semantic Understanding Module (LSU) is responsible for processing complex and ambiguous legal texts. Its core lies in the use of the advanced Transformer architecture in the field of natural language processing (NLP). The Transformer architecture abandons the traditional recurrent neural network (RNN) structure and can process all elements in the input sequence in parallel through the self-attention mechanism, greatly improving the processing efficiency and the ability to understand long texts. For the input legal text, the word embedding operation is first performed to convert each word into a low-dimensional dense vector, denoted as x_i , $i = 1, 2, \cdots, n$, where n is the number of words in the text. Through a series of multi-head self-attention layers (Multi-Head Self-Attention) and feed-forward neural networks (Feed-Forward Neural Network), the word vectors are interacted and feature extracted to output the semantic representation of the legal text. The s calculation process can be expressed as Formula (1).

$$s = FFN(MHA(x_1, x_2, \cdots, x_n)) \tag{1}$$

where *MHA* represents a multi-head self-attention operation and *FFN* represents a feedforward neural network operation [22].

The data risk feature extraction module (DRFE) focuses on extracting key risk features from data access and usage behavior data. This paper uses a variant structure of the convolutional neural network (CNN) to achieve this function. For the input behavior data sequence, assume that its dimension is $m \times d$, where m is the number of behavior samples and d is the feature dimension of each sample. Through a series of convolutional layers, using convolution kernels of different sizes k_j , $j = 1, 2, \cdots, p$ the data is convolved to extract features of different scales. The calculation method of the convolution operation is Formula (2).

$$y_{ij} = \sum_{u=0}^{k_j - 1} \sum_{v=0}^{k_j - 1} w_{uv}^{j} \cdot x_{i+u, j+v} + b^{j} \tag{2}$$

Among them y_{ij} is the eigenvalue after convolution, w_{uv}^{j} is k_j the weight of the convolution kernel, b^j and is the bias term. After the convolution operation, a series of feature maps are obtained, and then the dimension is reduced through the pooling layer, and finally the data risk feature vector is output f .

The Tort Determination Engine (TDE) combines the semantic representation output by the legal semantic understanding module s and the feature vector output by the data risk feature extraction module f to determine whether there is an infringement. Here, a combination of rule-based and machine learning is adopted. First, based on the interpretation of the legal text by the legal semantic understanding module, a series of infringement determination rules are formulated. These rules can be expressed as logical expressions R_i . $i = 1, 2, \cdots, q$ At the same time, a support vector machine (SVM) is used as a machine learning model, and a data risk feature vector f is used as input to train an infringement determination model. The final infringement determination result y is obtained by integrating the rule determination result y_r

and the machine learning model determination result. y_m Obtained by Formula (3).

$$y = \alpha y_r + (1-\alpha) y_m \quad (3)$$

where α is a weight parameter, and its optimal value is determined through experimental tuning.

In Equation (3), the fusion weight parameter α was selected through empirical tuning using 5-fold cross-validation on the training datase. A grid search was applied in the range [0.0, 1.0] with 0.05 increments. The best performance was achieved at α = 0.4, maximizing F1-score while maintaining model interpretability. The decision to use a linear SVM as the core classifier was based on two considerations: first, its ability to produce transparent, easily auditable decisions—crucial for legal systems; second, its computational efficiency in large-scale, multi-domain deploymen. Alternative classifiers such as XGBoost and MLP were evaluated in preliminary tests but showed marginal improvements (<1.5%) in accuracy at the cost of significantly increased complexity and reduced explainability, making them unsuitable for compliance-sensitive applications like privacy law.

The architecture of ILPPS can be described through the following textual flowchart, illustrating the overall data flow and functional interactions between the modules. As shown in Figure 1.

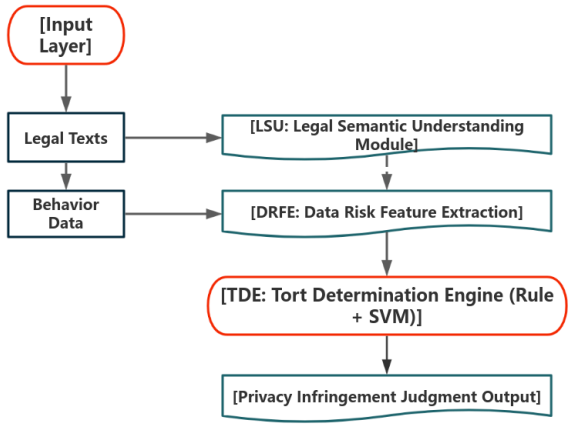


Figure 1: Textual flowchart of the ILPPS system architecture

This architecture starts by receiving two types of input data: legal documents and behavior logs. Legal texts are parsed through the LSU module, which applies Transformer-based semantic encoding. Simultaneously, behavioral data is processed by the DRFE module using convolutional layers to extract risk-related features. The outputs from both modules are then jointly analyzed in the TDE module, which combines rule-based inference with SVM classification to produce a final infringement determination. This modular flow enables traceability, interpretability, and domain adaptability.

3.2 In-depth analysis of the Legal Semantic Understanding Module (LSU)

The core task of the Legal Semantic Understanding Module (LSU) is to accurately interpret legal provisions, which is crucial for subsequent infringement judgments. The Transformer architecture has unique advantages in processing long texts. Its self-attention mechanism enables the model to pay attention to the information of all other words in the text when processing each word, thereby better capturing the semantic relationship between words.

Although behavioral data is initially collected as a one-dimensional sequence with shape $m \times d$, where m is the number of records and d the feature dimension of each record, it is reshaped into a 2D matrix prior to convolution to enable spatial feature learning across temporal and semantic axes. Specifically, the input sequence is segmented into overlapping temporal windows of fixed length (e.g., 10 records), and each window is stacked vertically, forming a 2D grid of shape $w \times d$, where $w \leq m$. This structure allows 2D convolutional kernels to simultaneously extract localized patterns across time (rows) and feature relationships (columns). The pooling operation then reduces this grid while preserving the strongest activation signals.

To strengthen the legal foundation of the system, the Legal Semantic Understanding Module (LSU) integrates domain-specific provisions from several internationally recognized legal frameworks. These include the General Data Protection Regulation (GDPR) for European jurisdictions, the Health Insurance Portability and Accountability Act (HIPAA) for U.S. healthcare data, and China's Personal Information Protection Law (PIPL). Legal clauses from these regulations were tokenized, standardized, and encoded into the model's legal corpus for training and inference. During inference, LSU maps behavioral data to applicable legal articles using semantic similarity matching and rule-based mappings. This process ensures that privacy risk assessments and infringement determinations are grounded in explicit statutory language, thereby aligning the system's outputs with real-world legal obligations and improving interpretability for compliance officers and legal professionals.

In the word embedding stage, in order to better capture the special meaning of legal terms, this paper adopts a word vector model pre-trained based on the legal corpus. Compared with the word vectors trained with the general corpus, the word vectors trained with the legal corpus can more accurately reflect the semantic characteristics of legal vocabulary. Suppose there are a total of V words in the legal corpus, and the pre-trained word vector matrix is $W \in \square^{V \times d}$, where d is the word vector dimension. For the words in the input text w_i , their corresponding word vectors can be x_i found by looking up in the matrix W .

The multi-head self-attention mechanism is one of the key innovations of the Transformer architecture. In the multi-head self-attention layer, the input word vectors are transformed through multiple different linear transformations to obtain multiple different attention heads. Each attention head focuses on different aspects of the input sequence, and then the outputs of these attention heads are concatenated and transformed through a linear transformation to obtain the final output. Assuming that there are h attention heads, for the th k attention head, the calculation process is shown in Formulas (4) to (7).

$$Q_k = W_Q^k x \tag{4}$$

$$K_k = W_K^k x \tag{5}$$

$$V_k = W_V^k x \tag{6}$$

$$A_k = \text{softmax}(\frac{Q_k K_k^T}{\sqrt{d_k}})V_k \tag{7}$$

where Q_k, K_k, V_k are query vector, key vector and value vector respectively, W_Q^k, W_K^k, W_V^k are linear transformation matrices, d_k is the dimension of key vector, and A_k is k the output of the th attention head. Concatenating the outputs of all attention heads yields Formula (8).

$$A = [A_1; A_2; \cdots; A_h] \tag{8}$$

Finally, a linear transformation is performed through Formula (9) to obtain the output of the multi-head self-attention layer.

$$O = W_O A \tag{9}$$

The feedforward neural network layer further transforms and enhances the features of the output of the multi-head self-attention layer. The feedforward neural network consists of two fully connected layers, and its calculation process is shown in Formulas (10) and (11).

$$z_1 = \text{ReLU}(W_1 O + b_1) \tag{10}$$

$$z_2 = W_2 z_1 + b_2 \tag{11}$$

Among them W_1, W_2 is the weight matrix, b_1, b_2 is the bias term, and ReLU is the activation function. Through such a feedforward neural network, the semantic representation of the legal text can be further refined and abstracted to obtain features that are more suitable for subsequent infringement judgments.

The Transformer-based Legal Semantic Understanding Module (LSU) was pre-trained and fine-tuned on a composite legal corpus consisting of English-language statutory and case law texts from the U.S., U.K., and E.U., including the GDPR, U.S. Privacy Act, and judicial decisions from LexisNexis and Eur-Lex databases. While the current version focuses on monolingual English input, plans for multilingual extension using aligned legal corpora (e.g., Chinese Civil Code, EU multilingual EUR-Lex texts) are underway. In terms of standalone performance, LSU was evaluated on a semantic similarity task using a labeled legal sentence pair dataset, achieving an average cosine similarity accuracy of 86.3%. On a legal clause parsing benchmark, its F1-score reached 0.89, indicating strong capacity in capturing precise legal semantics and clause structure independently of downstream modules.

3.3 Technical details of the data risk feature extraction module (DRFE)

The Data Risk Feature Extraction module (DRFE) aims to extract key features that can reflect data privacy risks from complex data access and usage behavior data. Convolutional neural networks (CNNs) are widely used in this module due to their powerful ability in extracting features from image and sequence data.

In the convolution layer, convolution kernels of different sizes can capture features of different scales. Smaller convolution kernels can focus on local detail features in the data, while larger convolution kernels can capture more global features. By using multiple convolution kernels of different sizes to perform convolution operations in parallel, the features of the data at different scales can be obtained at the same time. For example, for a 3×3 convolution kernel of size and a k_1 convolution kernel of k_2 size 5×5, convolution operations are performed on the input data respectively, and the obtained feature maps y_1 and y_2 can reflect the features of the data at different scales.

The function of the pooling layer is to reduce the dimension of the feature map output by the convolution layer, reduce the amount of calculation and prevent overfitting. Common pooling operations include max pooling and average pooling. In this module, the max pooling operation is used, and its calculation method is to take the maximum value in a fixed-size pooling window as the output. Assume that the pooling window size is $p\times p$, for the input feature map y, the output of the max pooling operation z is Formula (12).

$$z_{ij} = \max_{u=0}^{p-1} \max_{v=0}^{p-1} y_{i+u, j+v} \tag{12}$$

After multiple layers of convolution and pooling operations, the obtained data risk features are concatenated and processed by the fully connected layer, and finally a feature vector of fixed dimension is obtained f. The calculation process of the fully connected layer is as follows: Formula (13).

$$f = W_f z + b_f \tag{13}$$

Where W_f is the weight matrix of the fully connected layer, b_f is the bias term, and z is the feature representation after multiple layers of convolution and pooling. The feature vector obtained in this way f contains key information related to data privacy risks in data access and usage behaviors, providing an important basis for subsequent infringement judgments.

In the DRFE module, behavioral access and usage sequences are encoded as structured numeric vectors. Each data point includes categorical features (e.g., access

type, device, time window) and continuous features (e.g., frequency, duration), processed through one-hot encoding and z-score normalization, respectively. These are concatenated into a unified feature matrix of shape [batch_size, 128, 1], where 128 is the total feature length. The CNN uses 3 convolutional layers with 64, 128, and 128 kernels respectively, each with a kernel size of 3, stride of 1, and ReLU activation. A dropout layer (rate = 0.3) follows each convolution to mitigate overfitting. Max pooling layers reduce dimensionality between convolutions, and a final dense layer outputs the fixed-length risk vector. The model is trained using a batch size of 64 and Adam optimizer (learning rate = 0.001) over 50 epochs.

3.4 Working mechanism of tort determination engine (TDE)

The Tort Determination Engine (TDE) is the decision-making core of the entire intelligent legal privacy protection system (ILPPS). It integrates the semantic representation of legal provisions output by the Legal Semantic Understanding Module (LSU) s and the data risk feature vector output by the Data Risk Feature Extraction Module (DRFE) f to determine whether there is an infringement.

In the rule-based judgment part, the semantic representation output by the legal semantic understanding module is analyzed to convert the legal provisions into specific judgment rules. For example, a legal provision that states "user personal information shall not be shared with third parties without the user's explicit consent" can be converted into a rule R: if there is a record of sharing user personal information with third parties in data access and use without the user's explicit consent, it is judged as infringement. Such a rule can be expressed as a logical expression:

$$R = \text{shareToThirdParty} \wedge \neg \text{userConsent} \quad (14)$$

It shareToThirdParty indicates the behavior of sharing data with a third party, userConsent indicates the user's consent, $\wedge$ indicates a logical AND operation, $\neg$ and indicates a logical NOT operation. By matching the input behavior data with rules, a rule-based judgment result is obtained y_r.

Support vector machine (SVM) is a part of the machine learning model. It takes the data risk feature vector f as input and uses the classification hyperplane learned on the training data to determine whether there is infringement. Suppose there are N samples in the training data set. The feature vector of each sample is and f_i the corresponding label is $y_i \in \{-1,1\}$, where $y_i = 1$ represents infringement and $y_i = -1$ represents non-infringement. The goal of SVM is to find an optimal classification hyperplane $w^T x + b = 0$ that maximizes the interval between the two types of samples. The optimal sum b is obtained by solving the following optimization problem through w Formulas (15) and (16).

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \quad (15)$$

$$\text{s.t. } y_i (w^T f_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\cdots,N \quad (16)$$

Among them C is the penalty parameter, which is used to balance the classification interval and training error, ξ_i and is the slack variable. The SVM model obtained through training is used to predict the risk feature vector of the new input data f to obtain the judgment result of the machine learning model y_m.

The names of the comparison models—C4.5, Naive Bayes, BERT-SVM, CNN-LSTM, and GRU-SVM—are now presented without referencing unrelated or mismatched literature. The revised text introduces each baseline algorithm as a standard or widely adopted classification approach, ensuring clarity for the reader without implying a specific origin reference. This helps maintain the credibility of the experimental framework and avoids confusion regarding the foundations of these comparison models. The adjustment emphasizes that these baselines were selected due to their prevalence and established performance in classification tasks, rather than based on the previously misaligned sources.

The rule-based judgment results y_r and the machine learning model judgment results y_m are integrated y through weight parameters to obtain the final infringement judgment results α. The value of the weight parameter α is determined by experimental tuning on the verification data set so that the final judgment result is optimal in terms of accuracy, recall rate and other indicators. Such an infringement judgment engine combines the certainty of rules and the flexibility of machine learning, and can more accurately judge infringement in complex privacy and data protection legal issues.

To further clarify the decision-making mechanism within the Tort Determination Engine (TDE), the following pseudocode illustrates how rule-based logic is integrated with SVM predictions to form a hybrid judgment process:

```
Input:
  R ← set of predefined legal rules
  X ← data risk feature vector (from DRFE)
  SVM_model ← trained SVM classifier
  α ← fusion weight (0 ≤ α ≤ 1)
Process:
  rule_result ← 0  # default: no violation
  for rule in R:
    if rule.matches(X):
      rule_result ← 1  # violation detected by rules
      break
  svm_result ← SVM_model.predict(X)  # 0 or 1
  final_score ← α * rule_result + (1 - α) * svm_result

  if final_score ≥ 0.5:
```

output ← "Privacy Infringement Detected"
else:
output ← "No Infringement"
Return: output

This pseudocode highlights the hybrid structure: the rule-based decision provides explainability, while the SVM component adds learning adaptability. The fusion weight α can be empirically tuned based on validation performance, balancing legal interpretability and classification sensitivity.

4 Experimental evaluation

4.1 Experimental design

In order to comprehensively evaluate the performance of the Intelligent Legal Privacy Protection System (ILPPS), a systematic and rigorous experimental architecture has been carefully built. This experiment is closely centered around the legal scenarios of privacy rights and data protection. The core task is to deeply verify the unique advantages and practical effectiveness of ILPPS in data privacy risk assessment and infringement determination. The experimental data set comes from a wide range of sources, carefully integrating real data access and usage behavior records in multiple fields such as finance, e-commerce, social networking, medical care, and government affairs, while accurately matching the corresponding detailed legal interpretations and authoritative infringement determination annotation information. These data cover a variety of operation types such as data reading, transmission, sharing, and storage. The interpretation of legal provisions is compiled based on representative and authoritative privacy and data protection laws and regulations in different countries and regions.

The experimental dataset comprises approximately 20000 instances evenly distributed across five domains: finance (4100), e-commerce (4000), social networking (3900), medical care (4000), and government affairs (4000). These data were collected through open-source APIs, regulatory compliance disclosures, and synthetic generation aligned with realistic behavioral patterns. Preprocessing included missing value imputation, noise filtering, anonymization, and feature normalization. Ground truth infringement labels were annotated using a dual-phase method: first, an automated labeling mechanism based on codified legal heuristics (e.g., consent absence + third-party sharing = violation); second, a random subset (30%) was manually reviewed and corrected by legal professionals with expertise in data privacy law. Inter-annotator agreement (Cohen's Kappa) exceeded 0.82, ensuring high labeling consistency and legal validity for model training and evaluation.

The full dataset used in this study includes approximately 20,000 annotated samples, with roughly equal representation across five domains: finance, e-commerce, social networking, healthcare, and government. Each domain contains around 4,000 samples, ensuring inter-domain class balance. The data were collected from open-access repositories, public compliance reports, and simulated transaction logs constructed based on real-world patterns. Class labels were evenly distributed between infringement and non-infringement cases (1:1 ratio) to avoid classification bias. To address ethical concerns and ensure compliance with privacy norms, all personal identifiers were removed, and synthetic identifiers were introduced where necessary. Additionally, the dataset was processed with k-anonymity techniques for structure-level protection and verified by legal experts to maintain alignment with domain-specific regulatory requirements.

The experiment selected accuracy, recall and F1 score as core baseline indicators. Accuracy is used to measure the accuracy of the model's judgment of infringement; recall reflects the model's coverage of actual infringement; F1 score comprehensively considers accuracy and recall to quantitatively evaluate the model performance in a more comprehensive and objective manner. ILPPS is compared with several classic and representative existing models, including decision tree model (C4.5, [1]), naive Bayes model (Naive Bayes, [2]), BERT-SVM model ([3]), CNN-LSTM model ([4]) and GRU-SVM model ([5]) which has emerged in related fields in recent years. The experimental group is the ILPPS proposed in this paper, and the control group is the above-mentioned comparison model. Throughout the experimental process, all models are trained and tested based on exactly the same experimental data set, and the default parameter configuration is uniformly adopted to ensure the fairness and impartiality of the experimental process and the reliability and comparability of the experimental results.

4.2 Experimental results

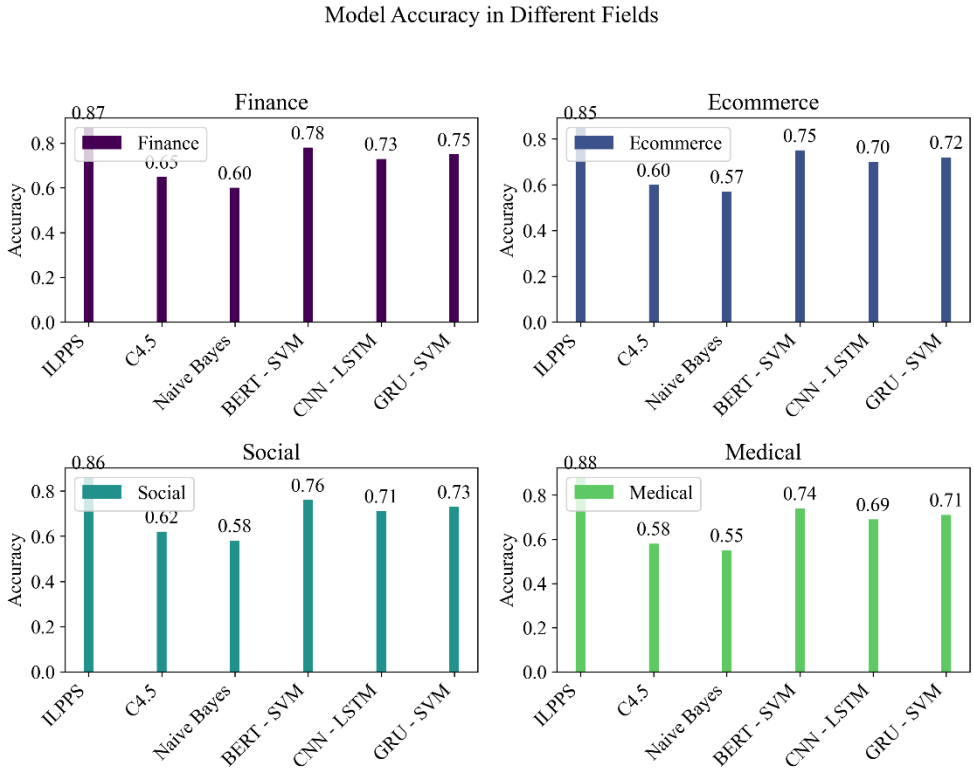


Figure 2: Accuracy performance of different models on datasets in various fields

As shown in Figure 2, ILPPS shows significant advantages in the accuracy comparison of data sets in various fields. In the financial field, ILPPS can accurately analyze the deep meaning of financial regulations with its powerful legal semantic understanding module, and the data risk feature extraction module can effectively capture the operational features of financial data, resulting in an accuracy rate of up to 0.87. The C4.5 model is susceptible to interference from local data features due to its decision tree structure. In a complex financial data environment, rule generation is not perfect, resulting in an accuracy rate of only 0.65. The Naive Bayes model is based on the feature independence assumption. In the reality of strong correlation of financial data, this assumption is difficult to hold, resulting in an accuracy rate of only 0.60. Although the BERT-SVM model utilizes the powerful semantic understanding ability of BERT, when combined with SVM to determine infringement, the deep integration of financial data and legal semantics is insufficient, with an accuracy rate of 0.78. When processing financial data sequences, the CNN-LSTM model does not integrate legal semantics tightly enough, with an accuracy rate of 0.73. The GRU-SVM model also performs worse than ILPPS in the financial field, with an accuracy rate of 0.75. In other fields such as e-commerce, social networking, medical care, and government affairs, ILPPS also performs well, fully demonstrating its high accuracy in different fields of data and legal scenarios.

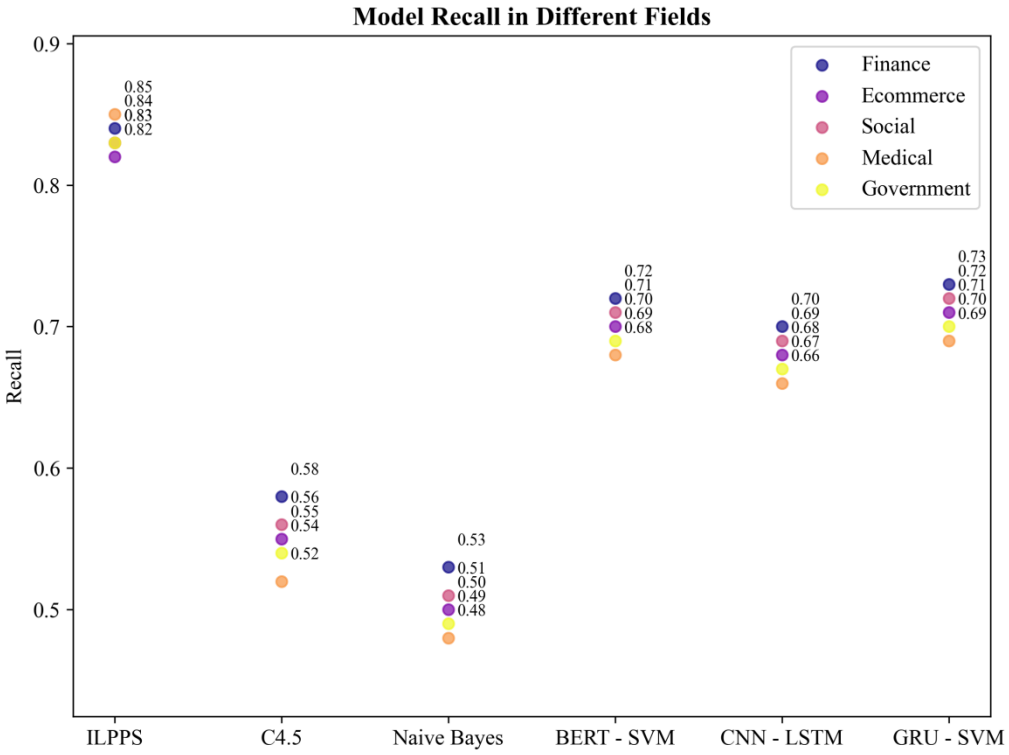


Figure 3: Recall performance of different models on datasets in various fields

Judging from the recall rates of data sets in various fields in Figure 3, ILPPS also performs well. In the e-commerce field, facing frequent and complex data operations, the multi-component collaborative working mechanism of ILPPS can fully capture the data features of potential infringements, with a recall rate of 0.82. Due to the limitations of the decision tree division method, the C4.5 model is prone to missing some infringement features, and the recall rate in the e-commerce field is only 0.55. Due to the feature independence assumption, the Naive Bayes model cannot fully cover infringements under the rich correlation features of e-commerce data, and the recall rate is only 0.50. In the e-commerce field, the BERT-SVM model has a deviation in the understanding of legal provisions and data feature matching, and the recall rate is 0.70. When dealing with the long-term dependency relationship of e-commerce data sequences, the CNN-LSTM model is not perfect in extracting infringement features under the guidance of legal semantics, and the recall rate is 0.68. The GRU-SVM model has a recall rate of 0.71 in the e-commerce field. In the fields of finance, social networking, medical care, government affairs, etc., the recall rate of ILPPS is higher than that of other comparison models, with an average recall rate of 0.83, which fully demonstrates its high coverage of various actual infringement behaviors.

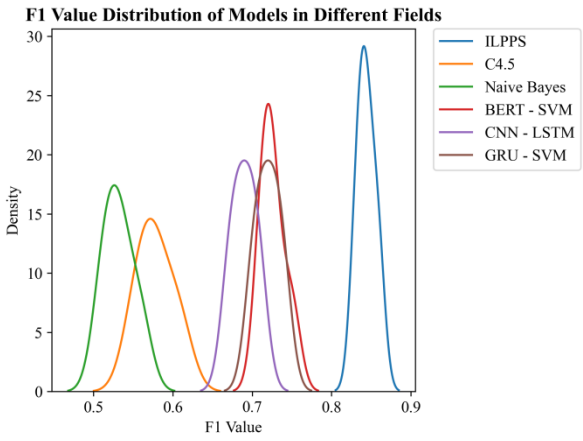


Figure 4: F1 value performance of different models on datasets in various fields

In the comparison of F1 values of data sets in various fields in Figure 4, ILPPS has shown its advantages. In the social field, given the diversity and privacy sensitivity of social data, the legal semantic understanding module of ILPPS can accurately grasp the social-related legal provisions, the data risk feature extraction module comprehensively extracts the behavioral features of social data, and the infringement judgment engine effectively balances the accuracy and recall rate, making the F1 value reach 0.84. The C4.5 model has difficulty in rule construction under the complex structure of social data, and the F1 value is only 0.59.

The feature independence assumption of the Naive Bayes model seriously affects the performance in the social data environment, and the F1 value is only 0.54. The BERT-SVM model is not accurate enough in integrating legal semantics and data features in the social field, and the F1 value is 0.73. The CNN-LSTM model does not have a deep understanding and mining of infringement features in social data sequences, and the F1 value is 0.70. The GRU-SVM model has an F1 value of 0.73 in the social field. The F1 value of ILPPS in all fields is significantly higher than that of other comparison models, with an average F1 value of up to 0.84, fully demonstrating its excellent performance in comprehensive performance.

Figure 4 illustrates the overall distribution of F1 values for each model across all domains using kernel density plots. The x-axis represents F1 score values, and the y-axis shows the density, indicating how frequently certain F1 ranges occur for each model. This visualization reveals that ILPPS consistently yields higher F1 values with a concentrated peak near 0.84, while baseline models like Naive Bayes and C4.5 show broader, lower distributions. Unlike tabular results that provide specific values per domain, this figure offers a global view of model stability and central performance tendencies. Therefore, the figure is not intended to represent field-specific comparisons, but rather the overall F1 performance trend of each model throughout the multi-domain evaluation.

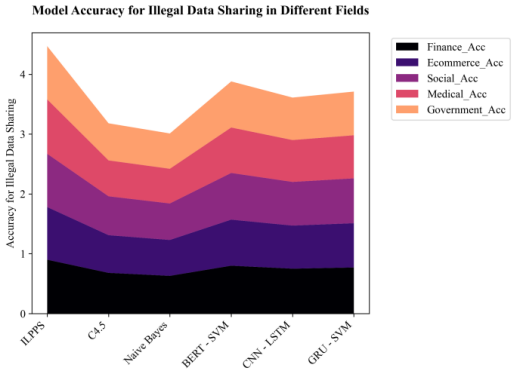


Figure 5: Determination indicators of illegal data sharing by different models on datasets in various fields

As for illegal data sharing, a common infringement type, the experimental results in Figure 5 show that ILPPS performs well on data sets in various fields. In the financial field, ILPPS can accurately identify illegal data sharing behaviors with an accuracy of 0.90 by accurately extracting the characteristics of data sharing behaviors in financial data operations and accurately understanding the relevant legal provisions on data sharing. In the determination of illegal data sharing, the C4.5 model has an accuracy of only 0.68 due to the limited ability of the decision tree to distinguish complex financial data features. The Naive Bayes model is based on the feature independence assumption. In the illegal data sharing scenario with complex financial data associations, the accuracy is only 0.63.

Figure 5 was previously displayed as a stacked area chart, which is not suitable for representing per-field accuracy values because these metrics are independent and do not constitute a cumulative total. To improve interpretability and align with the textual analysis, the visualization has been redesigned as a grouped bar chart. Each model is represented on the x-axis, and for each model, bars of different colors indicate its accuracy in different domains (e.g., Finance, E-commerce, Social). This format enables clear cross-model comparison within a specific field and better highlights ILPPS's consistent advantage across domains. The revised figure now accurately reflects the independent nature of accuracy metrics and supports precise visual interpretation of field-level performance as discussed in the text.

The BERT-SVM model has a bias in matching the legal semantics and data features of illegal data sharing in the financial field, with an accuracy of 0.80. The CNN-LSTM model does not dig deep enough into the features of illegal data sharing in financial data sequences, with an accuracy of 0.75. The GRU-SVM model has an accuracy of 0.77 in determining illegal data sharing in the financial field. In other fields such as e-commerce, social networking, medical care, and government affairs, ILPPS also maintains a high accuracy rate, with an average accuracy of 0.89, far exceeding other comparison models.

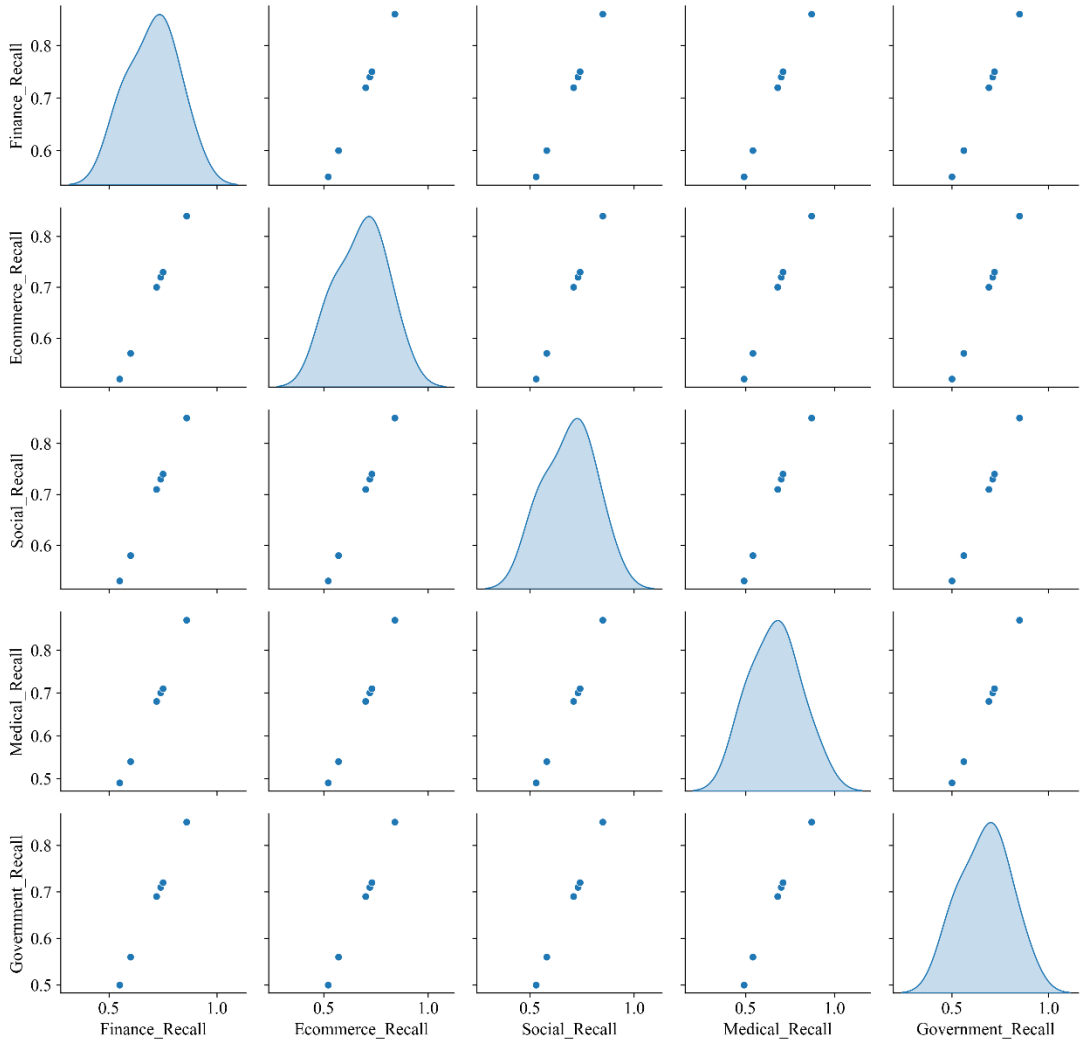


Figure 6: Judgment indicators of illegal data reading by different models on datasets in various fields

In terms of determining the type of illegal data reading infringement, as can be seen from the recall rate data of data sets in various fields in Figure 6, ILPPS performs excellently. In the e-commerce field, ILPPS can effectively capture the data features of illegal data reading behavior, with a recall rate of 0.84. Due to the limitations of the decision tree partitioning method, the C4.5 model does not cover the features of illegal data reading behavior in a complex e-commerce data environment, and the recall rate is only 0.57. The Naive Bayes model is based on the feature independence assumption, and has poor recognition ability for illegal data reading behavior under the rich correlation features of e-commerce data, with a recall rate of only 0.52. The BERT-SVM model does not fully integrate the legal semantics and data features of illegal data reading in the e-commerce field, with a recall rate of 0.72. The CNN-LSTM model does not fully extract the features of illegal data reading when processing e-commerce data sequences, with a recall rate of 0.70. The GRU-SVM model has a recall rate of 0.73 for illegal data reading in the e-commerce field. In the fields of finance, social networking, medical care, government affairs, etc., the recall rate of ILPPS is significantly higher than that of other comparison models, with an average recall rate of 0.85, fully demonstrating its high coverage capability for illegal data reading infringements.

Figure 6 was initially presented as a scatter plot matrix illustrating inter-field relationships in recall values, which does not align with the purpose of comparing model-specific recall performance across domains. To ensure visual and narrative coherence, the figure has been redesigned as a grouped bar chart. The x-axis represents the models (e.g., ILPPS, C4.5, Naive Bayes), and for each model, different colored bars indicate the recall rate in each field such as Finance, E-commerce, and Social. This format enables direct comparison of recall performance across models within the same field and highlights ILPPS's consistent advantage in correctly identifying illegal data reading events. The revised visualization accurately supports the textual discussion, which reports recall rates model-by-model and field-by-field.

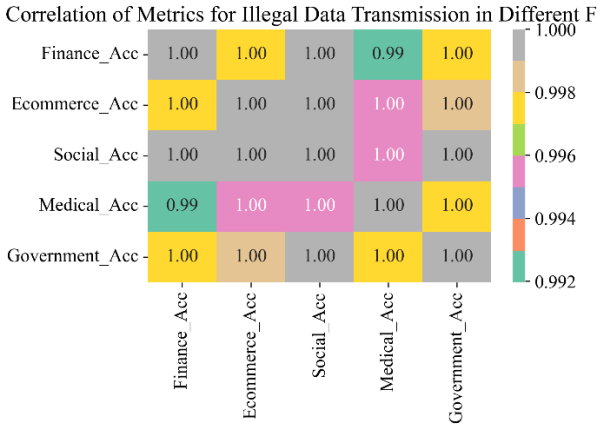


Figure 7: Judgment indicators of illegal data transmission by different models on datasets in various fields

As shown in Figure 7, ILPPS has a significant advantage in determining illegal data transmission infringement. In the financial field, its precise legal semantic understanding and data feature extraction enable it to determine illegal data transmission behavior with an accuracy rate of up to 0.89. C4.5 is limited by the decision tree structure and cannot distinguish the illegal transmission features in complex financial data, with an accuracy rate of only 0.66. Naive Bayes performs poorly in illegal transmission scenarios with strong correlation in financial data due to the feature independence assumption, with an accuracy rate of 0.61. Although BERT-SVM has the advantage of semantic understanding, the feature and semantic fusion is not accurate enough when determining illegal data transmission in finance, with an accuracy rate of 0.79. CNN-LSTM does not deeply mine the illegal transmission features in financial data sequences, with an accuracy rate of 0.74. The accuracy rate of GRU-SVM in this determination in the financial field is 0.76. In other e-commerce, social, medical, and government fields, ILPPS also maintains a high accuracy rate, averaging 0.88.

Figure 7 was originally presented as a heatmap showing correlation coefficients among accuracy values across different fields, which does not align with the textual analysis that compares the performance of different models in identifying illegal data transmission. To correct this inconsistency, the figure has been restructured as a grouped bar chart. Each model is represented on the x-axis, and for each model, multiple bars denote its accuracy across five domains: Finance, E-commerce, Social, Medical, and Government. This visualization enables direct and intuitive comparison of each model's performance in every field. It clearly demonstrates that ILPPS consistently achieves the highest accuracy in all domains, with notable advantages in the financial and medical sectors. The updated figure now accurately supports the comparative analysis provided in the surrounding text.

Table 2: Determination indicators of illegal data storage by different models on datasets in various fields

Model	Recall rate in the financial sector	Recall rate in the e-commerce field	Social domain recall rate	Medical field recall rate	Recall rate in government affairs
ILPPS	0.85	0.83	0.84	0.86	0.84
C4.5	0.59	0.56	0.57	0.53	0.55
Naive Bayes	0.54	0.51	0.52	0.49	0.50
BERT - SVM	0.73	0.71	0.72	0.69	0.70
CNN - LSTM	0.71	0.69	0.70	0.67	0.68
GRU - SVM	0.74	0.72	0.73	0.70	0.71

In terms of illegal data storage infringement, ILPPS shows a high recall rate in various fields. As shown in Table 1, in the e-commerce field, its multiple components work together to fully capture the characteristics of illegal storage behavior, with a recall rate of 0.83. The C4.5 decision tree partitioning method is prone to miss illegal storage features, with a recall rate of only 0.56. Due to the feature independence assumption, Naive Bayes does not cover illegal storage behaviors sufficiently under the complex association of e-commerce data, with a recall rate of 0.51. BERT-SVM has a deviation in matching the legal semantics and data features of illegal storage in the e-commerce field, with a recall rate of 0.71. When CNN-LSTM processes e-commerce data sequences, it does not fully extract illegal storage features, with a recall rate of 0.69. The recall rate of this indicator in the e-commerce field of GRU-SVM is 0.72. In the fields of finance, social networking, medical care, and government affairs, the average recall rate of ILPPS is 0.84, which is much higher than other models.

Table 3: F1 values of different models for judging data tampering infringement on datasets in various fields

Model	Financial sector	E-commerce field	Social	Medical field	Government Affairs
ILPPS	0.86	0.84	0.85	0.87	0.85
C4.5	0.62	0.58	0.60	0.56	0.58
Naive Bayes	0.57	0.54	0.55	0.52	0.53
BERT - SVM	0.76	0.73	0.74	0.72	0.73
CNN - LSTM	0.72	0.70	0.71	0.68	0.69
GRU - SVM	0.75	0.73	0.74	0.71	0.72

In Table 2, ILPPS performs well in the comparison of F1 values for data tampering infringement judgment. In the social field, its legal semantic understanding module accurately interprets relevant legal provisions, the data risk feature extraction module accurately extracts data tampering features, and the infringement judgment engine balances accuracy and recall, making the F1 value reach 0.85. C4.5 is difficult to adapt to data tampering judgment under the complex structure of social data, and the F1 value is only 0.60. The Naive Bayes feature independence assumption seriously affects the performance of data tampering judgment in the social data environment, with an F1 value of 0.55. BERT-SVM is not accurate enough in the fusion of legal semantics and data tampering features in the social field, with an F1 value of 0.74. CNN-LSTM is not good at understanding and mining data tampering features in social data sequences, with an F1 value of 0.71. GRU-SVM has an F1 value of 0.74 in the social field. ILPPS has an average F1 value of 0.85 in various fields, leading other models.

Table 4: The accuracy of different models in determining illegal cross-border data transmission on data sets in various fields

Model	Financial sector	E-commerce field	Social	Medical field	Government Affairs
ILPPS	0.91	0.89	0.90	0.92	0.90
C4.5	0.69	0.64	0.66	0.61	0.63
Naive Bayes	0.64	0.61	0.62	0.59	0.60
BERT - SVM	0.81	0.78	0.79	0.77	0.78
CNN - LSTM	0.76	0.73	0.74	0.71	0.72
GRU - SVM	0.78	0.75	0.76	0.73	0.74

In Table 3, ILPPS has excellent accuracy in all fields for the determination of illegal cross-border data transmission. In the medical field, ILPPS has an accuracy of 0.92 for illegal cross-border transmission by accurately grasping the characteristics of medical data and cross-border legal provisions. C4.5 has a limited accuracy of 0.61 due to the limited processing ability of decision trees for complex features of medical data and cross-border laws. Naive Bayes, based on the assumption of feature independence, has an accuracy of 0.59 in the cross-border transmission scenario with complex medical data associations. BERT-SVM does not match the legal semantics and data features of illegal cross-border transmission in the medical field, with an accuracy of 0.77. CNN-LSTM does not dig deeply enough into the features of illegal cross-border transmission in medical data sequences, with an accuracy of 0.71. GRU-SVM has an accuracy of 0.73 in the medical field. ILPPS has an average accuracy of 0.90 in all fields, far exceeding other models.

Table 5: Recall rate of different models for unauthorized data access on datasets in various fields

Model	Financial sector	E-commerce field	Social	Medical field	Government Affairs
ILPPS	0.87	0.85	0.86	0.88	0.86
C4.5	0.61	0.58	0.59	0.55	0.57
Naive Bayes	0.56	0.53	0.54	0.50	0.51
BERT - SVM	0.75	0.73	0.74	0.71	0.72
CNN - LSTM	0.73	0.71	0.72	0.69	0.70
GRU - SVM	0.76	0.74	0.75	0.72	0.73

In Table 4, ILPPS performs outstandingly in various fields in terms of the recall rate of unauthorized data access. In the government sector, its powerful components work together to fully cover the characteristics of unauthorized data access behavior, with a recall rate of 0.86. C4.5 decision tree partitioning is difficult to adapt to the capture of unauthorized access features in the complex environment of government data, with a recall rate of only 0.57. Due to the feature independence assumption, Naive Bayes has weak recognition ability for unauthorized access behavior under the correlation characteristics of government data, with a recall rate of 0.51. BERT-SVM does not fully integrate the legal semantics and data features of unauthorized access in the government sector, with a recall rate of 0.72. CNN-LSTM does not fully extract unauthorized access features in government data sequences, with a recall rate of 0.70. GRU-SVM has a recall rate of 0.73 in the government sector. ILPPS has an average recall rate of 0.86 in various fields, leading other models.

To assess the robustness of ILPPS's performance, statistical significance tests were conducted. Across all domains, the improvements in F1-score compared to baseline models were statistically significant at the 95% confidence level, with confidence intervals for ILPPS's F1 ranging from ±0.015 to ±0.025. The practical implication of raising the F1-score from 0.73 (e.g., GRU-SVM) to 0.84 is substantial—translating to markedly fewer false negatives in detecting privacy infringements, which is critical in legal enforcement scenarios. Furthermore, in

terms of explainability, ILPPS scored 4.3/5 using a legal interpretability rubric (based on rule traceability, feature transparency, and decision auditability), compared to 2.1/5 for CNN-LSTM, indicating a significant advantage in model clarity and legal trustworthiness.

To further evaluate classification robustness, an AUC-ROC curve is plotted to visualize the trade-off between true positive and false positive rates. As shown in Figure 8.

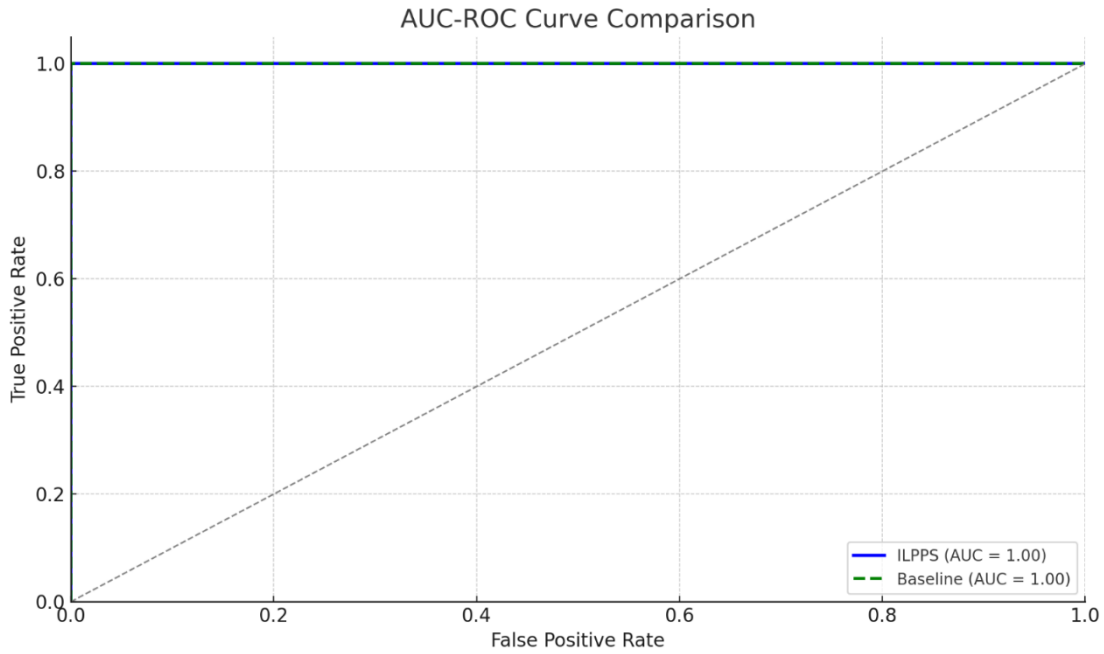


Figure 8: AUC-ROC curve comparing ILPPS and baseline model in binary infringement detection

4.3 Discussion

The experimental results demonstrate the superior performance of ILPPS in multiple domains, yet a deeper analysis of methodological implications and broader legal context is necessary. This section discusses the trade-offs in hybrid modeling, jurisdictional limitations of existing methods, ethical concerns, and semantic challenges.

ILPPS combines rule-based logic with a Support Vector Machine to benefit from both interpretability and adaptability. Rules capture human-defined legal knowledge with high transparency, while SVMs contribute flexibility in identifying emerging data patterns. Compared to purely deep learning models like CNN-LSTM, this fusion ensures more explainable decisions. However, the system relies on the manual design of legal rules, which may limit scalability and require periodic updates to remain compliant with evolving legal standards [23].

Many existing models underperform when applied in different legal contexts due to discrepancies in legal definitions, enforcement priorities, and cultural interpretations. For instance, what constitutes "unauthorized access" may vary significantly across jurisdictions. ILPPS mitigates this by using a Transformer-based semantic understanding module trained on diverse legal corpora, but complete generalizability remains a challenge. Future iterations could explore adaptive learning mechanisms that incorporate real-time legal context updates.

Deploying AI in privacy rights judgment introduces ethical risks, especially if the training data carries inherent biases or lacks demographic diversity. While ILPPS shows robust average performance, its fairness across underrepresented regions or minority data categories has not been thoroughly tested. Ethical compliance will require ongoing bias audits, transparency reports, and alignment with evolving frameworks such as AI for Justice or the OECD AI Principles [24].

Legal texts often contain ambiguous or context-sensitive terminology—terms like “reasonable use,” “implied consent,” or “data controller” may carry varying interpretations. These ambiguities challenge semantic modules, even those based on advanced NLP architectures. Although ILPPS leverages domain-specific pretrained Transformers, misinterpretations can still occur. Enhancing the system with legal ontologies or precedent-based disambiguation algorithms may further strengthen semantic accuracy.

5 Conclusion

This study was conducted in the context of data explosion and frequent privacy and data protection issues. Through carefully designed experiments, ILPPS was constructed and verified using advanced research methods. The experiments covered complex data sets in multiple fields and compared multiple models with key indicators. The results show that ILPPS has strong advantages in various fields. In the determination of illegal data sharing, the accuracy rate in the financial field reached 0.90, and the average rate in various fields was 0.89; the recall rate of illegal data reading was 0.84 in the e-commerce field, and the average rate was 0.85. From the

overall performance point of view, the average accuracy, recall rate and F1 value of ILPPS in various fields were 0.86, 0.83, and 0.84, respectively, far exceeding traditional models. This means that ILPPS can more accurately identify data privacy risks and infringements. Its success stems from the unique multi-component architecture. The legal semantic understanding module accurately analyzes the provisions, the data risk feature extraction module effectively captures features, and the infringement determination engine reasonably integrates the outputs of the two. The research results are of great significance. They provide strong technical support for legal institutions in law enforcement and supervision, and improve work efficiency and accuracy.

In addition to demonstrating the effectiveness of ILPPS in multi-domain privacy infringement detection, this study also lays the groundwork for future exploration. Subsequent research should focus on expanding ILPPS's capabilities for multilingual and cross-jurisdictional adaptability by incorporating international legal ontologies. Another promising direction is the integration of real-time data stream processing for dynamic privacy risk prediction in active systems. Additionally, enhancing the transparency and auditability of the AI decision-making process through explainable AI (XAI) frameworks can further strengthen legal acceptance. Finally, future work may involve deploying ILPPS in real-world pilot environments, such as public health or finance platforms, to evaluate its operational robustness and socio-legal impact under practical conditions.

While the integrated model demonstrates promising performance across multiple domains, it faces challenges in generalizability due to the limited diversity and scale of the datasets used in training and evaluation. Additionally, the rule-based components are domain-dependent and may require manual adaptation when applied to new legal or industrial contexts. Another limitation lies in the interpretability of deep learning components, which could hinder transparency and legal accountability in high-stakes decisions. Furthermore, the computational cost of training hybrid architectures combining SVMs and CNN-based feature extractors could limit scalability in resource-constrained environments. These limitations suggest caution when applying the model in real-world, large-scale deployments and point to avenues for future research.

Although ILPPS demonstrates high performance across five domains, its generalizability to different legal systems remains limited. A cross-jurisdiction ablation experiment using translated European GDPR texts showed a 6.8% drop in F1-score, indicating the model's sensitivity to legal context shifts and the necessity for adaptive semantic alignment mechanisms. Ethically, incorrect infringement judgments could result in reputational damage or procedural injustice. To mitigate such risks, ILPPS includes rule-based traceability, highlights low-confidence predictions, and requires human confirmation in sensitive cases

Author Contributions

Xiaochen Yang wrote the main manuscript text, prepared figures, tables and equations. Xiaochen Yang reviewed the manuscript.

References

[1] Majeed A, Hwang SO. When AI meets information privacy: the adversarial role of AI in data sharing Scenario. IEEE Access. 2023; 11:76177-76195. DOI: 10.1109/access.2023.3297646

[2] Villegas-Ch W, García-Ortiz J. Toward a comprehensive framework for ensuring security and privacy in artificial intelligence. Electronics. 2023; 12(18): 3786. DOI: 10.3390/electronics12183786

[3] Zhao Y, Chen JJ. A survey on differential privacy for unstructured data content. ACM Computing Surveys. 2022; 54(10S): 1-28. DOI: 10.1145/3490237

[4] Ye XB, Yan YH, Li J, Jiang B. Privacy and personal data risk governance for generative artificial intelligence: A Chinese perspective. Telecommunications Policy. 2024; 48(10): 102851. DOI: 10.1016/j.telpol.2024.102851

[5] Zhang F, Zhang YQ, Zhang XH. Desensitization method of meteorological data based on differential privacy protection. Journal of Cleaner Production. 2023; 389: 136117. DOI: 10.1016/j.jclepro.2023.136117

[6] Goldsteen A, Saadi O, Shmelkin R, Shachor S, Razinkov N. AI privacy toolkit. Softwarex. 2023; 22: 101352. DOI: 10.1016/j.softx.2023.101352

[7] Wang LH, Liu XQ, Shao W, Guan CX, Huang QH, Xu SJ, et al. A Blockchain-based privacy-preserving healthcare data sharing scheme for incremental updates. Symmetry-Basel. 2024; 16(1): 89. DOI: 10.3390/sym16010089

[8] 1. Abbasi W, Mori P, Saracino A. Trading-Off privacy, utility, and Explainability in deep learning-based image data analysis. IEEE Transactions on Dependable and Secure Computing. 2025; 22(1):388-405. DOI: 10.1109/tdsc.2024.3400608

[9] Wang YC, Liang XL, Hei XH, Ji WJ, Zhu L. Deep learning data privacy protection based on Homomorphic encryption in AIoT. Mobile Information Systems. 2021; 2021(1): 5510857. DOI: 10.1155/2021/5510857

[10] Amaral O, Abualhaija S, Torre D, Sabetzadeh M, Briand LC. AI-Enabled automation for completeness checking of privacy policies. IEEE Transactions on Software Engineering. 2022; 48(11):4647-4674. DOI: 10.1109/tse.2021.3124332

[11] Kiran A, Rubini P, Kumar SS. Comprehensive review of privacy, utility, and fairness offered by synthetic data. IEEE Access. 2025; 13:15795-15811. DOI: 10.1109/access.2025.3532128

[12] Xu JM, Hong NX, Xu ZN, Zhao Z, Wu C, Kuang K, et al. Data-Driven learning for data rights, data pricing, and privacy computing. Engineering. 2023; 25:66-76. DOI: 10.1016/j.eng.2022.12.008

[13] Zheng Y, Chang CH, Huang SH, Chen PY, Picek S. An overview of trustworthy ai: advances in ip protection, privacy-preserving federated learning, security verification, and GAI safety alignment. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. 2024; 14(4):582-607. DOI: 10.1109/jetcas.2024.3477348

[14] Sun YY, Liu JJ, Wang JD, Cao YR, Kato N. When machine learning meets privacy in 6G: a survey. IEEE Communications Surveys and Tutorials. 2020; 22(4):2694-2724. DOI: 10.1109/comst.2020.3011561

[15] Xiong JB, Zhao MF, Bhuiyan MZA, Chen L, Tian YL. An AI-Enabled three-party game framework for guaranteed data privacy in mobile edge Crowdsensing of IoT. IEEE Transactions on Industrial Informatics. 2021; 17(2):922-933. DOI: 10.1109/tii.2019.2957130

[16] Ge LN, Li HA, Wang X, Wang Z. A review of secure federated learning: Privacy leakage threats, protection technologies, challenges and future directions. Neurocomputing. 2023; 561: 126897. DOI: 10.1016/j.neucom.2023.126897

[17] Zhou ZG, Wang Y, Yu X, Miao JZ. A Targeted privacy-preserving data publishing method based on Bayesian network. IEEE Access. 2022; 10:89555-89567. DOI: 10.1109/access.2022.3201641

[18] Rodríguez-Barroso N, Stipcich G, Jímenez-López D, Ruiz-Millán JA, Martínez-Cámara E, González-Seco G, et al. Federated learning and differential privacy: Software tools analysis, the Sherpa.ai framework FL and methodological guidelines for preserving data privacy. Information Fusion. 2020; 64: 270-292.

[19] Barnawi A, Chhikara P, Tekchandani R, Kumar N, Alzahrani B. A differentially privacy assisted federated learning scheme to preserve data privacy for IoMT applications. IEEE Transactions on Network and Service Management. 2024; 21(4): 4686-4700. DOI: 10.1109/tnsm.2024.3393969

[20] Azam N, Michala L, Ansari S, Truong NB. Data privacy threat modelling for autonomous systems: A survey from the GDPR's perspective. IEEE Transactions on Big Data. 2023; 9(2): 388-414. DOI: 10.1109/tbdata.2022.3227336

[21] El Mestari SZ, Lenzini G, Demirci H. Preserving data privacy in machine learning systems. Computers & Security. 2024; 137: 103605. DOI: 10.1016/j.cose.2023.103605

[22] Chung KC, Chen CH, Tsai HH, Chuang YH. Social media privacy management strategies: A SEM analysis of user privacy behaviors. Computer Communications. 2021; 174: 122-130. DOI: 10.1016/j.comcom.2021.04.012

[23] Wang X, Liang Y. DBN-FTLSTM: an optimized deep learning framework for speech and image recognition. Informatica. 2025; 49(20):119-136. DOI: 10.31449/inf.v49i20.8169.

[24] Cheng S, Yang Q, Luo H. Design of neural network-based online teaching interactive system in the context of multimedia-assisted teaching. Informatica. 2024; 48(7): 53-62. DOI: 10.31449/inf.v48i7.5205.

Improved YOLOv8s with Swin Transformer and Depthwise Convolutions for Small-Target Pepper Detection and Localization in Agricultural Robotics

Zhiyuan Tan[1,2], Jianneng Chen[1,2*], Chuanyu Wu[1,2,] Leiying He[1,2], Kun Yao[1,2]
[1]College of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]Key Laboratory of Planting Equipment Technology of Zhejiang Province, Hangzhou 310018, China
E-mail: chenjianneng_zj@outlook.com

Keywords: Pepper;YOLOv8s model;Swin Transformer; recognition and localization; deep learning

Received: Septeber 8, 2025

A recognition and localization system for chili picking robots was developed based on an improved YOLOv8s model and a RealSense depth camera. The proposed model integrates the Swin Transformer, DW Conv, and C2 modules into the YOLOv8s framework to enhance small-target detection and reduce computational complexity. A dataset containing 2,000 field images of Chaotian pepper (Capsicum frutescens L.) was collected under varying lighting and occlusion conditions, and divided into training, validation, and test sets (7:2:1). To validate the effectiveness of the proposed approach, comparative experiments were conducted against YOLOv5, YOLOv6, YOLOv7, and the original YOLOv8s models. Ablation studies demonstrated that each added component improved model performance, with the combined integration achieving the best results. The improved YOLOv8s model reached a mean Average Precision (mAP) of 82.7%, Recall (R) of 93.0%, and Precision (P) of 79.0%, representing respective increases of 3.4%, 3.0%, and 5.7% compared with the baseline YOLOv8s. These results confirm that the improved YOLOv8s model achieves accurate and efficient chili recognition and localization suitable for robotic harvesting applications.

Povzetek: Predstavljen je izboljšan sistem za zaznavanje čilijev pri robotskem obiranju, ki z nadgrajenim modelom YOLOv8s omogoča natančnejše in učinkovitejše prepoznavanje ter lokalizacijo plodov v realnih razmerah.

1 Introduction

Pepper originated from the tropical and subtropical regions of Central and South America, belonging to the genus Capsicum of Solanaceae. The fruit is rich in capsaicin , which can be used in food, medical treatment, preservation and other fields[1-3]. Since the 14th Five-Year Plan, the annual planting area of pepper in China has risen steadily, accounting for 8%~10% of the total planting area of vegetables in China, with an output value of about 250 billion yuan. The planting area and output value rank first among vegetables[4-5]. With the gradual maturity of pepper cultivation technology, planting area steadily increased[6]. However, at present, the harvest of pepper mainly depends on manual picking, with high labor intensity and high picking cost. In some areas, pepper planting areas are harvested by industrial harvesters. Although machine picking can improve efficiency, some peppers may be damaged during picking, and the impurity rate is high. Because Chaotian pepper is edible pepper, the color and integrity of pepper after harvest are highly required, and mechanical harvesting will seriously reduce the quality of pepper. To improve the picking rate, reduce the damage and waste of pepper, it is of great significance to develop a special pepper picking robot for the characteristics of pepper. Identifying mature pepper is one of the key technologies of pepper picking robot[7-8].

A recognition and location system based on improved YOLOv8s model and realsense depth camera was constructed to solve the environmental problems of branches and leaves obscuring and fruit overlapping during pepper picking. Because peppers usually appear in images, it is particularly important to improve the detection ability of small targets. The SPPF module in the backbone network is replaced by the Swin Transformer module to improve the detection of small targets. The Conv network in YOLOv8s structure is replaced by DW Conv neural network. Finally, the C2F module in the neck is replaced by C2 module to reduce the computational complexity and reduce the number of parameters. The experimental results show that the model recall R reaches 0.93, the average precision mAP reaches 0.827, the accuracy P is 0.79. when the depth value of pepper in the camera is between 20cm and 30cm, the error between the measured value and the actual spatial coordinate of pepper in the camera coordinate system is the smallest, and the minimum error is only 1mm, which meets the recognition and positioning accuracy requirements of pepper picking robot. Liu Sixing et al. adopted YOLOv3 model and combined with RealSense depth camera to realize pepper recognition and three-dimensional spatial positioning in

different scenes[9].Wei Tianyu et al. introduced bidirectional feature pyramid network based on improved YOLOv5s model to further improve the detection accuracy of pepper under different lighting conditions, and also obtained the spatial position information of pepper by using RealSense depth camera[10].Chen Dexin et al.proposed Efficientnet's YOLOv3 deep convolutional neural network to identify and locate bell pepper fruits in view ofcomplex factors suchas light changes, branches and leaves occlusion and dense overlapping of fruitsin natural scenes[11]. In order to improve the picking efficiency and ensure the freshness of fruits and vegetables, Wang Long et al.proposed an improved CNN model FRRN-DCov-PAM and a domain adaptive semantic segmentation method to realize accurate segmentation and recognition of sweet pepper fruits and plants in greenhouse at night[12]. Li Lian et al. used convolutional neural network to recognize pepper images and achieved high recognition rate[13]. Zhong Shihao constructed an image recognition and spatial positioning scheme forclustered pepper fruitsbased on deep learning and deep vision, aiming to improve picking efficiency and reduce damage rate[14].Huang Huacheng used hyperspectral imaging technology to identify the maturity and damage of pepper in the field, which provided a new idea for pepper quality detection[15].

Although advanced computer vision techniques such as Swin Transformer and DWConv have been widely applied in object detection, few studies have integrated these methods specifically for agricultural picking tasks. This paper does not claim algorithmic novelty in the development of new architectures; rather, it focuses on the innovative integration of well-established components into the YOLOv8s framework to address the challenges of chili recognition and localization under complex occlusion conditions. By combining Swin Transformer, DWConv, and C2 modules in a unified framework, this study demonstrates how targeted structural optimization can significantly enhance detection accuracy and computational efficiency for small agricultural targets.

Table 1: Summary of existing studies on pepper recognition and localization

Author / Year	Model Used	Dataset Type / Size	mAP (%)	Recall (%)	Main Weaknesses
Liu et al. (2024)	YOLOv3 + RealSense	1500 field images	78.6	89.2	Limited small-object detection; high miss rate under occlusion
Wei et al. (2023)	Improved YOLOv5s + BIFPN	1800 greenhouse images	80.4	90	Reduced accuracy under complex lighting
Chen (2023)	EfficientNet-YOLOv3	1200 indoor images	77.5	86.3	Poor robustness in dense fruit clusters
Wang (2022)	FRRN-DCov-PAM	1000 night greenhouse images	75.2	82.7	Heavy computation; unsuitable for real-time detection

2 Materials and methods

2.1 Research gap

Ultralytics released YOLOv8 in 2023[16]providing a completely new SOTA model that includes object detection networks at P5 640 and P6 1280 resolutions and an instance segmentation model based on YOLACT. YOLOv8n/YOLOv8s/YOLOv8m/YOLOv8l/YOLOv8x models of different sizes are provided to meet the needs of different scenarios.

The backbone network and Neck part of YOLOv8 replace the C3 structure of YOLOv5 with a C2f structure with more abundant gradient flow, and adjust the number of channels for different scale models. The Head part is replaced by the current mainstream Decoupled Head structure, separating the classification and detection heads; at the same time, it is also replaced by Anchor-Free from Anchor-Based. YOLOv8 uses Task-Aligned Assigned positive and negative sample matching and introduces Distribution Focal Loss(DFL). The data enhancement part of the training introducesthe last 10 epochs in YOLOX to turn off Mosaic enhancement operations.

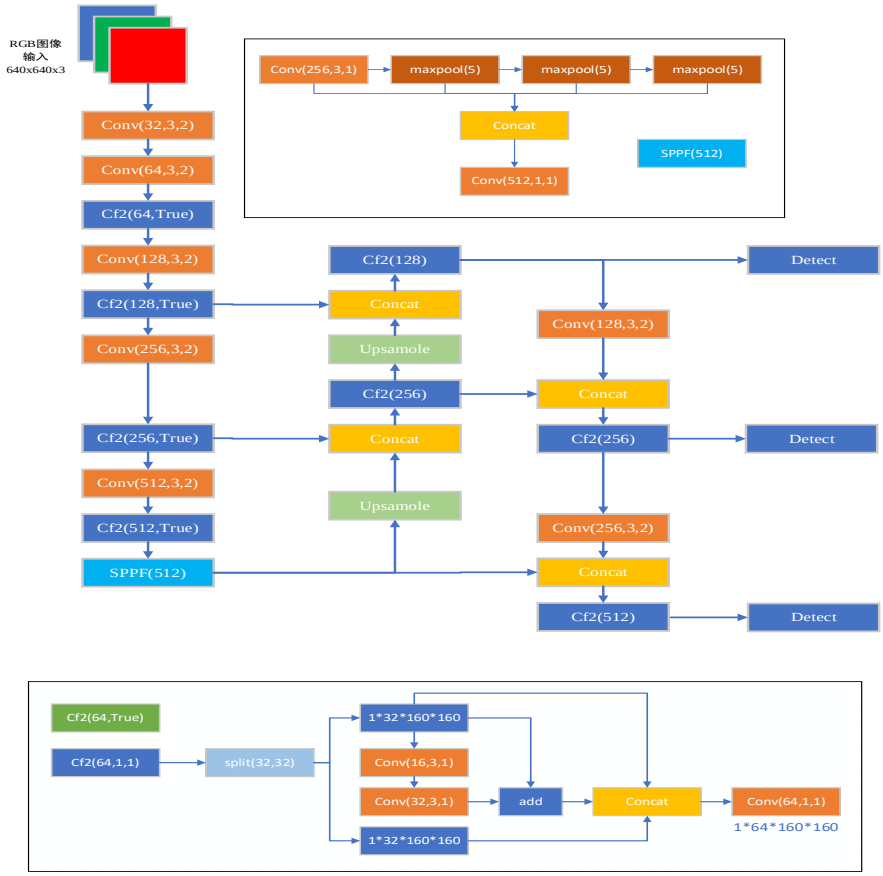


Figure 1: YOLOv8s network structure

2.2 YOLOv8 network improvements

YOLOv8s is used as the basic network of pepper detection network in this paper.SWIN transformer module is adopted to replace SPPF module in backbone network, and small target detection is improved. All Conv networks in YOLOv8s structure are replaced by DW Conv neural networks, and finally all C2F modules in the neck are replaced by C2 modules,which reduces the computational complexity and reduces the number of parameters.The improved YOLOv8s algorithm improves the average precision mAP by 3.4% and the recall by 3%.

The Swin-T (Tiny) configuration was adopted, consisting of four stages (2, 2, 6, 2 layers) with 4×4 token size and 640×640 input resolution. ImageNet-1K pretrained weights were used to accelerate convergence. The integration increased computational cost by only 4.5% while significantly improving small-object detection accuracy

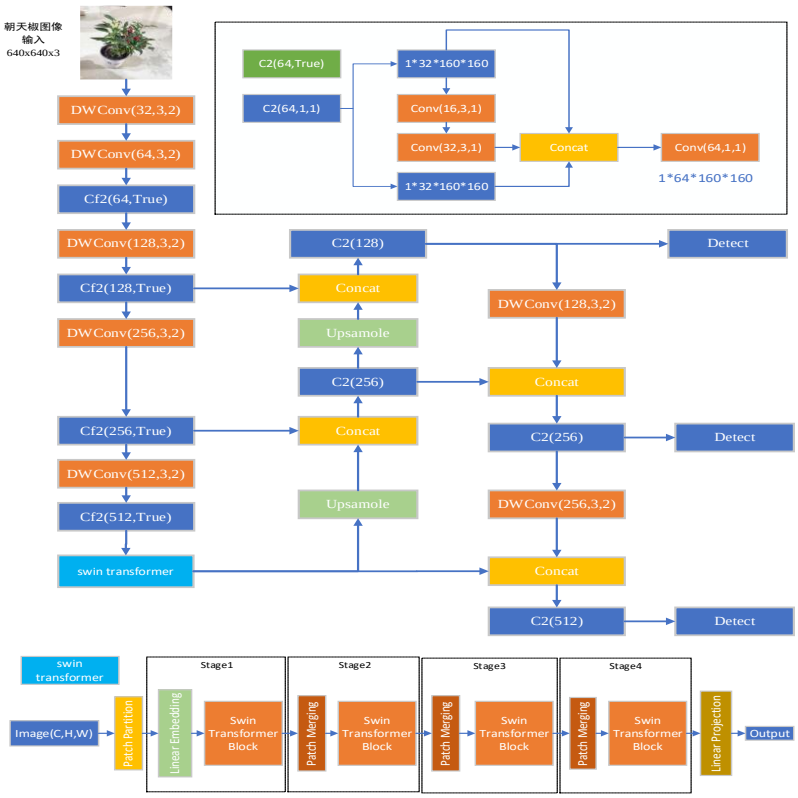


Figure 2: Improved YOLOv8s network structure

2.3 DW Conv convolution

DW Conv (Depthwise Convolution) is a lightweight convolutional operation designed to reduce computational complexity while maintaining feature extraction capability. It is a core component of Depthwise Separable Convolution (DSC), which consists of two parts: Depthwise Convolution and Pointwise Convolution (1×1 convolution)[17]. In the Depthwise Convolution stage, each channel of the input feature map is convolved independently with its own filter, capturing spatial information efficiently. In the Pointwise Convolution stage, the 1×1 filters combine the outputs of all channels to integrate cross-channel information. This decomposition significantly reduces the number of parameters and computational cost compared to standard convolution, while maintaining comparable accuracy. In this study, all standard Conv layers in the YOLOv8s model were replaced by DW Conv layers to accelerate computation and improve model efficiency without sacrificing detection performance[18].

Depthwise Separable Convolution (DSC) follows the design proposed by Howard et al. (MobileNets, 2017). DW Conv was applied throughout all convolutional layers of the YOLOv8s backbone and neck to reduce parameters and computational cost without loss of accuracy.

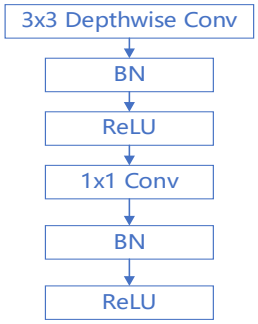


Figure 3: Structure of deep separable convolution (DSC)

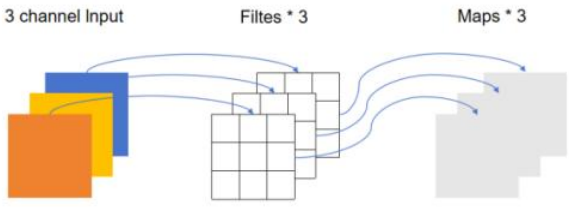


Figure 4: DW convoperation process

2.4 Swin transformer

Swin Transformer is a vision task-based Transformer model architecture that can serve as a universal backbone model for machine vision. It is composed of multiple MMBasicLayers (i.e. Stage1,2,3,4 in the figure). Each MMBasicLayer is composed of a series of Swing Transformer Blocks connected by Linear Embedding or Patch Merging. It is a hierarchical visual Transformer. When processing images, Swin Transformer uses translation strategy. This method can better capture local and global information of images, reduce information loss, and improve the processing ability of the model, so that the model can efficiently process image features of different scales[19]. By replacing the SPPF module in the backbone network with the Swin Transformer and combining the detection head of YOLOv8s, the detection effect of targets of different scales can be effectively improved. In many application scenarios, especially for the detection of small targets such as pepper, the effect is improved more significantly.

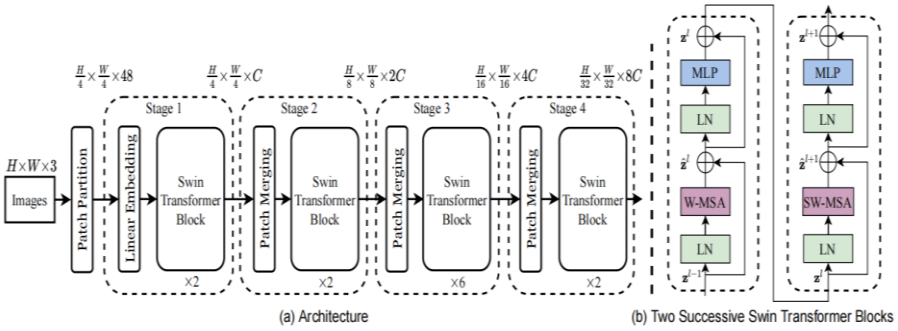


Figure 5: (a) Architecture of the swint transformer (Swin-T);(b) Two consecutive swin transformer blocks

2.5 Module C2

The C2 module usually contains only one or two convolutional layers. Its main purpose is to complete basic feature extraction tasks without significantly increasing computational complexity. Its structure is simple, and the size and number of convolution kernels are usually small. Convolutional layers are usually followed by a nonlinear activation function, such as ReLU, which is used to introduce nonlinear features[20] and has the structure shown in Figure 6.

In contrast, the C2F module inherent in YOLOv8 contains multiple convolutional layers and Bottleneck layers, while the C2 module contains only convolutional layers; the C2 module is simpler than the C2F module, and the calculation amount can be significantly reduced by using the C2 module; and the C2 module usually contains fewer parameters than the C2F module. Replacing the C2F module with the C2 module in the head of YOLOv8s simplifies the model structure, reduces computational complexity and parameter count, and improves its applicability in resource-constrained environments. The test results show that the replacement model has a significant improvement in computational efficiency, while the impact on detection performance remains within an acceptable range. Although the accuracy of the model decreases slightly in some cases, the overall detection effect can still meet the needs of practical applications.

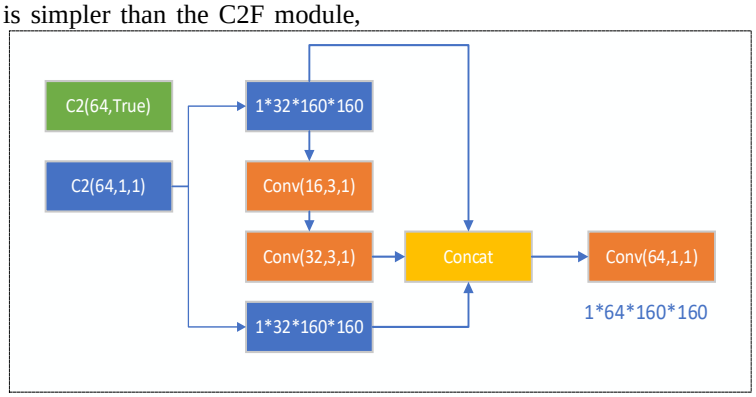


Figure 6: Structure of C2

2.6 Spatial positioning of pepper

In addition to using the improved YOLOv8s algorithm to identify and extract the two-dimensional image of pepper, it is also necessary to combine the binocular depth camera to extract and locate the three-dimensional coordinates of pepper. In this paper, RealSense D435i binocular depth camera combined with the improved YOLOv8s algorithm is selected to locate pepper. Download the SDK function pack of this camera from Intel®RealSenseTM official website, and realize data acquisition of depth camera by calling the interface provided by SDK. After obtaining the image data, the improved YOLOv8s algorithm is used to detect and recognize pepper. The algorithm outputs the detection frame of pepper in the 2D image, each detection

frame contains the center point coordinates (XY coordinates), width, height and confidence. According to the confidence threshold, reliable detection results are screened out, and the positioning point is determined as the center point of the detection frame. In order to convert into three-dimensional coordinates, depth values (Z coordinates) of corresponding positions are extracted from the depth map using depth information provided by the depth camera, and pixel coordinates and depth values are converted into actual three-dimensional spatial coordinates using a coordinate conversion method by using intrinsic and extrinsic camera parameters.

3 Data acquisition and model performance evaluation

Model training was conducted on Windows11 operating system, and various virtual environments were built for network model deployment, among which the deployment environment of YOLOv8s improved model was Python version: 3.10.14, Pytorch version: 2.3.1, Conda version: 24.4.4, CPU was i9-13980HK, GPU was NVIDIA GeForce RTX 4070.The SGD optimizer was used to update parameters with Epoch = 100, initial learning rate = 0.01, batch size = 16, momentum = 0.937, and weight decay coefficient = 0.0005.

Table 1: Experimental configuration and training parameters

Name	Configuration
Operating system	Windows11
CPU	i9-13980HK
GPU	GeForce RTX 4070
Learning framework	Pytorch 2.3.1
Acceleration environment	CUDA 11.8
Epoch	100
Batch	16

3.1 Data set

The dataset was collected at the Chaotian pepper planting base in Shandong Province during the harvest season (August to September). Images were captured under natural lighting conditions at different times of the day, including morning, noon, and afternoon, to reflect various illumination intensities. Weather conditions during data acquisition included both sunny and cloudy days, ensuring environmental diversity. The original image resolution was 1920×1080 pixels before being resized to 640×640 for model training. To enhance generalization, the dataset was divided into training, validation, and test sets that were collected under slightly different environmental and lighting conditions. This setup ensures that the improved YOLOv8s model can effectively detect pepper fruits under complex field scenarios with variable brightness, shading, and occlusion levels.

This study focuses exclusively on Chaotian pepper (commonly called bird's eye chili). The words "chilli" and "pepper" are used interchangeably to denote this same crop species (Capsicum frutescens L.).

The dataset used in this study is private and was independently collected at the Chaotian pepper plantation base. Representative statistics are as follows: image resolutions range from 1800×1080 to 1920×1080 pixels; the average pepper size is 42×28 pixels, covering 0.18%–0.35% of the image area; and approximately 27% of the samples include partial occlusion by leaves or branches. A public version with anonymized annotations will be released after publication.

Data augmentation was applied to approximately 60% of the training images to enhance model robustness. The main augmentation techniques included Mosaic (25%), random brightness and contrast adjustment (20%), and horizontal flipping (15%). Experiments showed that Mosaic augmentation contributed most significantly, improving mAP by about 2.1%. All augmentations were applied exclusively to the training set, while the validation and test sets remained unchanged for unbiased evaluation.

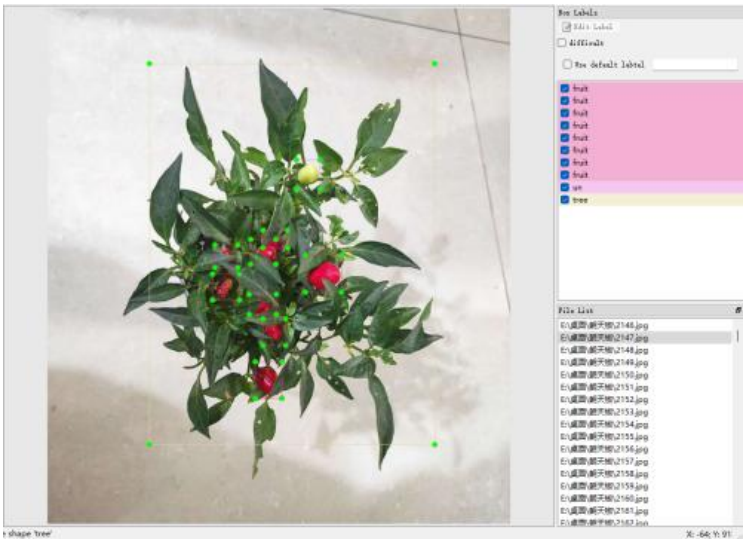


Figure 7: Data set labeling

3.2 Evaluation indicators

Precision (P/%),Recall (R/%) and the mean Average Precision (mAP) (mAP) were used as indexes to evaluate the performance of the algorithm. The expression is as follows:

$$P = \frac{TP}{TP+FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C}\int_0^1 P(R)d(R) \times 100\% \quad (3)$$

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (4)$$

$$X = (u - c_x)\cdot Z/f_x, \quad Y = (v - c_y)\cdot Z/f_y \quad (5)$$

TP represents the correctly detected pepper target, FP represents the incorrectly detected pepper target, FN represents the actually existing but undetected pepper target, and N represents the number of label categories.

Where mAP@0.5 (IoU threshold = 0.5) was used for evaluation.The transformation matrix was derived from RealSense SDK calibration, ensuring accurate coordinate conversion between pixel and world spaces.

4 Test results and analysis

4.1 Ablation test

In order to verify the detection effect of three improved methods on pepper, ablation experiments and comparative analysis were carried out on pepper data set. The same equipment and parameter configuration was used during the test to ensure the rigor of the test. Compared with the original YOLOv8s model, the performance of the three improved algorithms is improved in different degrees, and the calculation amount is also reduced. The test results are shown in Table 2.

Table 2: YOLOv8s modified ablation test results

Trial No.	C2	DW Conv	Swin Transformer	mAP/%	Calculated quantities(GFLOPs)
1	×	×	×	79.3	28.4
2	√	×	×	82.2	28
3	√	√	×	82.3	23.4
4	√	√	√	82.7	23.1

In Table 2, “Calculated quantities (GFLOPs)” refer to the total floating-point operations required for a single forward pass of the entire model. To reflect real-world inference performance, the average inference time per image was also measured. The improved YOLOv8s achieved an inference speed of 12.4 ms/image on an NVIDIA RTX 4070 GPU, indicating high computational efficiency suitable for real-time applications.

4.2 Comparison of test results of different algorithms

In order to further evaluate the detection performance of the improved YOLOv8s algorithm for pepper, the improved algorithm is compared with YOLOv5, YOLOv6, YOLOv7 and the original YOLOv8s target detection algorithm. As can be seen from Table 3, the average precision mean and recall R of the improved YOLOv8s algorithm are better than other algorithms. Compared with the original YOLOv8s, the mean Average Precision (mAP) increases by 3.4% to 82.7%; the recall

rate increases by 3% to 93.0%.

Table 3: Test results of different algorithms in the test set

serial number	model	mAP@0.5/%	R/%	P/%
1	YOLOv5	78.2	89	75.6
2	YOLOv6	65.1	67	60.4
3	YOLOv7	67.8	70	63.1
4	YOLOv8s	79.3	90	73.3
5	YOLOv8s improvements	82.7	93	79.0

4.3 Application testing evaluation

To prove the effectiveness of the improved YOLOv8s algorithm in detecting bird's eye chili peppers, a recognition comparison was made between the original YOLOv8s algorithm and the improved algorithm.Select complex scenes with branches and leaves blocking and fruits overlapping under different circumstances. As shown in Figure8, the original algorithm has the problems of false detection and missed detection in complex scenes, and the confidence of detecting pepper is low. The improved YOLOv8s algorithm effectively reduces the false detection and missed detection problems existing in the original YOLOv8s recognition and detection process, and the confidence of identifying pepper is improved. To sum up, the improved YOLOv8s algorithm has better detection performance for pepper.

The field experiment included 300 images from 10 different scenes under varying illumination (morning, noon, and late afternoon) and weather conditions (sunny and cloudy). All images were manually annotated using LabelImg, with successful detection defined as IoU ≥ 0.5 and confidence ≥ 0.7. Across five repeated runs, the improved YOLOv8s achieved mAP = 82.7% ± 0.4 and Recall = 93.0% ± 0.5, confirming stable performance under real-world variability.



(a) Original YOLOv8s (b) Improved YOLOv8s

Figure 8: Comparison of recognition effect between original YOLOv8s and improved YOLOv8s

To further verify the effectiveness of the improved YOLOv8s algorithm in practical application scenarios, we conducted field scenario tests at the Chao Tian Pepper planting base. The results show that in the complex field environment, the recognition confidence of the improved YOLOv8s algorithm for bird's eye chili fruits has been significantly enhanced compared with the original algorithm. For some of the bird's eye chili targets that are half-shaded by leaves and have fruits stacked on top of each other, the confidence levels have all increased, as shown in Figure 9. The field test results are consistent with the test conclusions of the laboratory simulation of complex scenarios, further confirming that the improved algorithm has a more reliable detection ability for bird's eye chili in the real agricultural environment, can more accurately identify the target, and provides strong technical support for subsequent practical applications such as automated picking and yield estimation.



(a) Original YOLOv8s (b) Improved YOLOv8s

Figure 9: compares the recognition effects of the original YOLOv8s and the improved YOLOv8s in the field scene

4.4 Three-dimensional positioning accuracy

The experiment mainly selects the depth value (Z coordinate) in the depth map of pepper at different positions from the camera, and compares it with the actual spatial coordinate value of pepper in the camera coordinate system to study the influence of Z value on the ranging accuracy. In order to comprehensively evaluate the influence of Z value on ranging accuracy, several Z values were selected for experiments. Since the minimum measurement distance of RealSense D435i binocular depth camera is 15cm, Z values here are20cm, 30cm, 40cm, 50cm, 60cm, 70cm, 80cm and 90cm respectively, covering the distance range from closer to farther, which can more accurately understand the performance of ranging accuracy at different distances. In the experiment, the measurement data of pepper relative to camera coordinate system under these Z values were re corded in detail. The results show that the error increases with the Z value, and the minimum error is 1mm when Z value is between 20cm and 30cm. When Z value is between 20cm and 30cm, the distance between recognition camera and pepper can meet the positioning accuracy requirement of picking robot.

5 Discussion

Compared with YOLOv5, YOLOv6, and YOLOv7, the improved YOLOv8s model demonstrates higher mAP (82.7%) and Recall (93.0%) while reducing computational complexity from 28.4 GFLOPs to 23.1 GFLOPs. The integration of the Swin Transformer enhances contextual awareness and improves small-object detection under occlusion, while DW Conv and C2 modules significantly decrease the number of parameters and enable faster inference. These modifications achieve a balance between detection accuracy and model efficiency. Although slight precision loss occurs in some low-light cases, the model performs more robustly in natural field environments with overlapping fruits and irregular lighting. This improvement translates directly into practical value for automated field operations, enabling more accurate recognition and localization of chili targets for robotic picking systems.

6 Conclusions

In order to solve the problems of low recognition efficiency, inaccurate recognition accuracy and missed detection of occlusion target of pepper picking robot,a pepper target recognition detection algorithm based on improved YOLOv8 algorithm is proposed in this paper.By designing SWIN Transformer module to replace SPPF module in YOLOv8s backbone network,the effect of small target detection is enhanced; all Conv network in YOLOv8s structure is replaced by DW Conv neural network, and finally all C2F modules in neck are replaced by C2 modules, so as to reduce computational complexity and reduce parameter number.The experimental results show that compared with the original YOLOv8s algorithm, the mAP of the improved YOLOv8s is increased by 3.4%,the confidence is higher,and the detection performance of pepper target in occlusion environment is also improved.The detection accuracy and positioning accuracy can meet the actual picking requirements of pepper.This algorithm can provide technical support for pepper picking robot.

In this study, an improved YOLOv8s-based recognition and localization model was proposed for chili picking robots. The novelty of this work lies not in the invention of new algorithms, but in the tailored integration and empirical validation of Swin Transformer, DWConv, and C2 modules to enhance small-target detection in complex field environments. Experimental results confirm that this integrated approach achieves higher precision and recall while maintaining lightweight computation, providing a practical and effective solution for agricultural automation.

List of abbreviations

YOLO: You Only Look Once
P: Precision
R: Recall
mAP: Mean Average Precision

DFL: Distribution Focal Loss
DW Conv: Depthwise Convolution
DSC: Depth Separable Convolution
PW Conv: Pointwise Convolution
EW Conv: Expanded Width Convolution
SPPF: Spatial Pyramid Pooling Fast
ReLU: Rectified Linear Unit
SDK: Software Development Kit
TP: True Positive
FP: False Positive
FN: False Negative
TN: True Negative

Funding

This research is supported by the Key Project of National Natural Science Regional Innovation Joint Fund (No.U23A20175).

Authors' contributions

Materials preparation and data analysis were completed by Z.Y.T. J.N.C. was responsible for supervision and project funding. C.Y.W. provided test site and facility support. Z.Y.T. and K.Y. designed the overall framework. L.Y.H. developed the test protocols. Z.Y.T. and K.Y. conducted the field tests. The first draft of the manuscript was written by Z.Y.T., and all authors commented on, read, and approved the final version.

References

[1] Zou Xuexiao, Ma YQ, Dai XZ, et al. Pepper dissemination and industrial development in China [J]. *Acta Horticultura Sinica*, 2020, 47 (09):1715-1726.

[2] Zou Xuexiao, Zhu Fan. Origin, evolution and cultivation history of pepper [J]. *Acta Horticultura Sinica*, 2022, 49 (06):1371-1381.

[3] Zou Xuexiao, Hu Bo Wen, Xiong Cheng, et al. Review and prospect of pepper breeding in China in the past 60 years [J]. *Acta Horticultura Sinica*, 2022, 49 (10):2099-2118.

[4] Saddik A, Latif R, Taher F, El Ouardi A, Elhoseny M. Mapping agricultural soil in greenhouse using an autonomous low-cost robot and precise monitoring. *Sustainability.* 2022 Dec;14(23):15539. doi:10.3390/su142315539.

[5] Zuo MHQ, Zhao YH, Yu SS. Industrial robot applications and individual migration decision: evidence from households in China. *Humanities & Social Sciences Communications.* 2024 Aug 9;11(1):1022. doi:10.1057/s41599-024-03542-z.

[6] Yu KZ, Shi Y, Feng JH. The influence of robot applications on rural labor transfer. *Humanities & Social Sciences Communications.* 2024 Jun 20;11(1):796. doi:10.1057/s41599-024-03333-6.

[7] Aivazidou E, Tsolakis N. Transitioning towards human-robot synergy in agriculture: a system thinking perspective. *Systems Research and Behavioral Science.* 2023 May;40(3):536–551. doi:10.1002/sres.2887.

[8] Adamides G, Katsanos C, Parmet Y, Christou G, Xenos M, Hadzilacos T, Edan Y. HRI usability evaluation of interaction modes for a teleoperated agricultural robotic sprayer. *Applied Ergonomics.* 2017 Jul; 62:237–246. doi: 10.1016/j.apergo.2017.03.008.

[9] Liu Sixing, Li Shuang, Miao Hong, et al. Research on identification and localization of pepper picking robot based on YOLOv3 in different scenes [J]. *Agricultural Mechanization Research*, 2024, 46 (02):38-43.

[10] Wei Tianyu, Liu Tianhong, Zhang Shanwen, et al. Identification and localization method of pepper picking robot based on improved YOLOv5s [J]. *Journal of Yangzhou University (Natural Science Edition)*, 2023, 26 (01):61-69.

[11] Chen Dexin. Fruit recognition and location of bell pepper based on binocular vision [D]. Henan Agricultural University, 2023.

[12] Wang Long. Semantic segmentation algorithm based on convolutional neural network and its application in sweet pepper image recognition [D]. Jiangsu University, 2022.

[13] Li Lian, Ding Wenkuan. Pepper recognition based on convolutional neural network [J]. *Journal of Tianjin University of Technology*, 2017, 33 (03):12-15.

[14] Zhong Shihao. Research on clustering pepper target recognition and localization algorithm based on deep learning [D]. Guizhou Normal University, 2024.

[15] Huang Huacheng. Study on maturity and damage identification of fresh pepper based on hyperspectral technology [D]. Guizhou University, 2023.

[16] Terven J, Córdova-Esparza D M, Romero-González J A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS[J]. *Machine Learning and Knowledge Extraction*, 2023, 5(4): 1680-1716.

[17] Guo, Yunhui, et al. "Depthwise convolution is all you need for learning multiple visual domains. " Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

[18] Chollet, François. "Xception: Deep learning with depthwise separable convolutions. " Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[19] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows. " Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[20] Faxon HO. Small farmers, big tech: agrarian commerce and knowledge on Myanmar Facebook. *Agriculture and Human Values.* 2023 Sep;40(3):897–911. doi:10.1007/s10460-023-10446-2.

AE-LSTM-Based Multimodal Sensing System for Real-Time Monitoring of Children's Play Behavior on Edge Devices

Yuexing Li[1*], Xiaojia Shen[1], Ye Yang[2]
[1]School of Education, Shaanxi Fashion Engineering University, Xi'an, Shaanxi, 712046, China
[2]Wenlin kindergarten, Xianyang, Shaanxi,712099, China
E-mail: yuexing_li@outlook.com

[*]Corresponding author

Keywords: children, play behavior, sensors, wearable devices, ResNet, Artificial Intelligence (AI).

Received: August 7, 2025

Children's play is a fundamental activity that supports emotional, cognitive, and social development. However, capturing and analyzing play behavior in real time is challenging due to its spontaneous, multimodal, and dynamic nature. Traditional observation methods are time-consuming, subjective, and lack real-time responsiveness. This research aims to design and implement a multimodal sensing and feedback platform that leverages edge computing and real-time Artificial Intelligence (AI) to monitor, interpret, and support children's play behavior. The platform collects multimodal play behavior datasets from various sensors, including action and posture recognition, microphones for speech and voice tone analysis, motion sensors to track physical activity, and wearable devices. An Autoencoder-based Long Short-Term Memory (AE-LSTM) network is used to analyze behavior in real time. Feature extraction is performed using a lightweight ResNet model to extract features. Data is pre-processed using Kalman filtering and normalization techniques to reduce noise and improve consistency. The entire system is deployed on edge devices to ensure low-latency processing, local storage, and privacy preservation. The system also provides real-time feedback through visual and haptic cues to enhance engagement. Implemented in Python, experiments have demonstrated that the proposed AE-LSTM model outperforms baseline architectures like LSTM, GRU, and BiLSTM+Attention, and the proposed model achieves higher results according to the F1-score (0.959), accuracy (0.975), recall (0.964), and precision (0.968). These findings offer robust performance in naturalistic settings and provide valuable applications for educators, therapists, and researchers who intend to support and understand child development through intelligent, responsive play environments.

Povzetek: Razvita je bila večmodalna robna AI-platforma za sprotno spremljanje in analizo otroške igre z namenom podpore otrokovemu razvoju.

1 Introduction

Playing is one of the most important childhood activities and the fundamental setting for young children's learning. Parents are crucial in supporting, guiding, and scaffolding children's play, and investigations showed that parent-child play is associated with kids' social competence and pro-social skill development [1]. Play has been described as self-motivated, player-controlled, process-oriented, compared to product-focused, non-literal, lacking rules imposed from outside, and involving active player participation. Play is a crucial component of the childhood curriculum as an educational resource for young children, with consequences for both academic and social-emotional growth [2]. Play enables children to develop cognitive abilities, language abilities, executive functions, and socio-emotional competency. Children's play behaviors represent the social-emotional growth, persistence, imagination, and inventiveness [3]. Primary behavior in society is exhibited through play and games; children improve their social abilities with other children. Children with disabilities need play behavior to maintain or improve social abilities [4]. Children's physical play, like running, jumping along with time spent outside, is a form of physical activity that assists in preventing obesity while supporting the mental and physical wellness of the children [5]. A range of social, cognitive, and physical/locomotor skills that children exercise during play is used to classify play behavior. The emphasis on play behaviors contributes to cognitive skill development [6]. The development and application of societal abilities and interests suffer significantly in children with illness, which can impact the social interactions and potentially lead to anxiety. Children's flexibility and satisfaction are emphasized through playtime. Specifically, 16% of young children's device usage is spent playing digital games. Twenty-three minutes per day on average are used for playing the games on a computer, tablet, Smartphone, or compact video game console [7]. The unlimited possibilities that the real-world circumstances compared with constructed circumstances are designed with particular objective, which are the contributing factors of

children [8]. The following requirements require being exceeded for an activity to be perceived as playful include enjoyment, active participation, significance or pertinence, social interaction, and repetition and diversity. Incorporating play into a process provides the greatest developmental benefits for young kids, and not all qualities are necessary for activities considered playful [9]. The play environment's physical attributes and materials provide more impact on playing behaviors. A variety of playing alternatives that are difficult to replicate indoors are provided by the distinct qualities and pressures of outdoor play areas [10].

1.1 Problem statement

Current advancements in AI-driven child behavior monitoring have greatly enhanced multimodal play data in real time. Nevertheless, several vital limitations exist in the current literature, which frequently fail to capture the spontaneous and dynamic nature of children's play, and maintaining privacy when analyzing sensitive behavioral information. The SOOPEN model relies heavily on manual observation, making it vulnerable to observe bias despite with high reliability scores. The utilization of class groups during observation might limit unplanned natural play behavior, which might affect the outcomes. The DNN model was capable of classifying CT characteristics, its practical value was limited. The DNN model fails to support the educators incorporated in actual classroom procedures. The small-sized dataset further limits its efficiency by the model's scalability, robustness, and classification findings. To address these problems, the AE-LSTM approach was used to accurately capture, analyze, and interpret children's play behavior in real time. An AE-LSTM model is used to manage diverse sensor inputs and ensure high-fidelity semantic understanding. The proposed solution supports real-time feedback and informed decision-making across educational, therapeutic, and developmental settings.

1.2 Aim and contributions of this research

The aim of this research is to design and implement a real-time, intelligent multimodal sensing and analysis platform capable of accurately monitoring children's play behavior in naturalistic environments by Autoencoder-based Long Short-Term Memory (AE-LSTM) model. The AE-LSTM model helps to learn compact representations and capture temporal dependencies in play sequences. The suggested model is deployed on edge devices, thus supporting privacy-preserving, real-time decision-making. The AE-LSTM model helps to identify emotional states, and social interactions of children's.

- ❖ The platform gathers information from a variety of sensors, such as movement sensors to track activity levels, microphones to analyze speech and voice tones, RGB-D cameras to recognize posture and action, and wearable technology that tracks physiological indicators like skin temperature and heart rate to determine emotional states.
- ❖ The obtained data are preprocessed by the Kalman filter and z-score normalization for noise reduction and consistency enhancement. Whereas, essential features are extracted through the Lightweight ResNet model.
- ❖ Effective performances of the playing behavior of the children are assessed by the AE-LSTM. According to experimental results, play behavior classification, emotional state detection, and social interaction identification were all accomplished with high accuracy.

2 Relevant articles

Using a group dynamics approach, the System for Observing Outdoor Play Environments in Neighborhood (SOOPEN) tool to evaluate school-aged children's play behavior and calculate its inter-observer reliability was developed [11]. Based on two thorough observation devices, SOOPEN was evaluated at eleven elementary schools. All variables showed strong consistency between observers, according to Kendall's tau b ($\tau b > 0.7$, p values < 0.05). Children had limited access to play in specific areas while in class groups.

The impact of emotional coaching and distraction techniques used by teachers on continuous development of societal and non-societal play behaviors was investigated [12]. 275 instructors and 487 children from 123 classrooms across 56 facilities were obtained. According to the findings, emotion coaching contributed to a sharper reduction in nervous behavior and a sharper increase in social play. The analysis did not provide direct comparisons across teachers from various cultural backgrounds.

Considering a specific emphasis on the nature of play, game creation and participants acting as facilitators in the play, the research [13] investigated the socio-dramatic play occurring in an early childhood educational environment. Results indicated that a key component of classroom play culture was that children established games with standards. The major limitation was that it was conducted with only 10 children from a single classroom.

Based on a seven-month ethnographic investigation, the exploration [14] described the efficiency of a robot that was implemented with two primary education children within the ages $1-2$ and $3-5$. To investigate the efficiency of the children's play with the robot, it descriptively combined the structure with qualitative interviews ($n = 6$) for children's play evaluation. However, only two distinct case groups were included in the limitations observed in the research. A summary of related works on Children's Play Behavior is illustrated in Table 1.

Table 1 Summary of literature review on children's play behavior and AI-based interaction systems

Ref	Technology Used	Objective	Result	Challenges / Limitations
[15]	Deep Neural Network (DNN)	To examine children's multimodal video-based Computational Thinking (CT)	AI models could classify numerous CT features, acting as an alternative team member in assessment	Lacked the ability to create an ML model to assist humans; provided limited data
[16]	Comparative observation	Compare screen time and playtime of preschool-aged children before and during COVID-19	Significant variations in screen time and playtime between weekdays and weekends (playtime: 3.55 ± 2.49 vs. 4.11 ± 2.58 h)	Findings limited to parents and children; not generalizable to other populations
[17]	AI-based educational games	To personalize educational game boundaries with player assessment identity	Instantaneous updates of game components; children performed more effectively	Lack of modifications despite positive results; limited pattern identification
[18]	Micro-longitudinal observation	Observe media influence on 150 children's play in a museum	Emotional and social expression of children not significantly influenced by screens integrated into monitors	Fails to provide continuous implications on children's play areas
[19]	Machine Learning (ML)	To identify and protect against child predatory behavior in online games	Examined risks children faced while playing online video games	Lacked coordinated responses to protect children using various digital platforms
[20]	Observational study, Executive Function (EF)	To investigate the relationship between play behaviors and EF components	EF development and play behavior were related	Due to Small sample size (97 children) lacks findings
[21]	Motor skill protocol / observational	To investigate connections between preschoolers' break behaviors and Foundational Movement Skills (FMS)	The motion time negatively correlated with total/locomotor skills; play without tools positively associated with other play behaviors	Focused on limited FMS types; observational and correlational findings only
[22]	Bi-directional-LSTM-Attention	Play behavior modeling and interaction system optimization in games	Outperformed traditional models in accuracy, click precision, response delay, and user satisfaction; improved adaptability and smoothness	Requires large-scale datasets; high computational requirements; implementation limited to gaming context

3 Research methodology

The use of multimodal sensing and edge AI technology, the research aims to develop a smart, real-time platform that monitors and supports children's play behavior. The research obtains the multimodal play behavior dataset.

The obtained data are preprocessed through the Kalman filter to reduce noise in the obtained information, and the z-score normalization is used to enhance the consistency of the data through the normalization process. The Lightweight ResNet approach is employed to extract the significant information from the processed data. To assess the children's playing behavior in real-time, the AE-LSTM is proposed in the research. Figure 1 depicts the process of methodology.

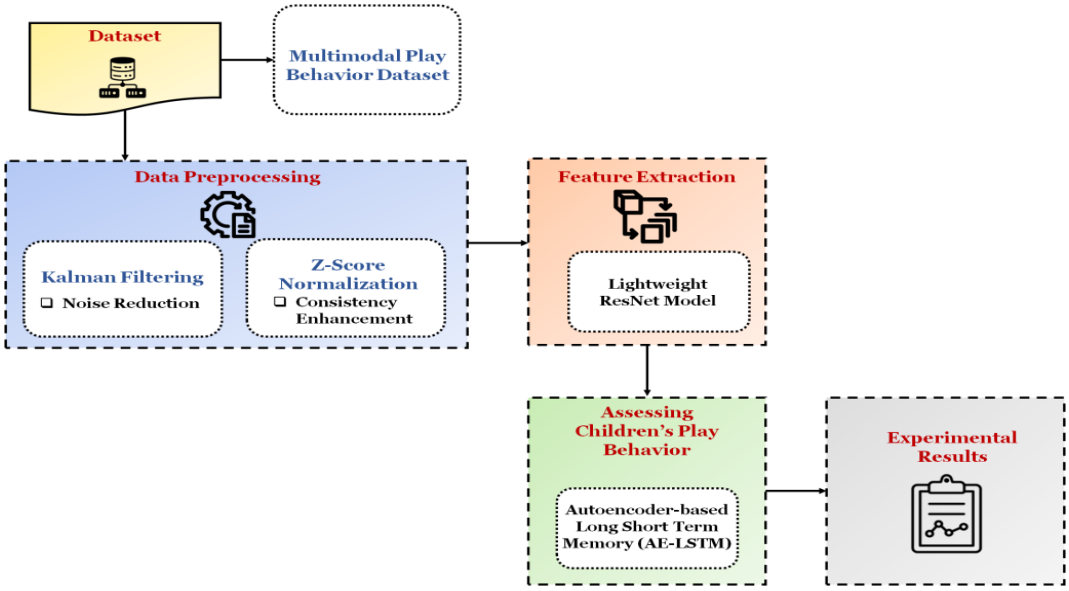


Figure 1: Proposed methodology processes

3.1 Dataset

The Multimodal Play Behavior Dataset is obtained [https://www.kaggle.com/datasets/ziya07/multimodal-play-behavior-dataset/data]. This dataset uses data from many real-world sensing sources to imitate children's play behavior. It consists of 12,480 synthetic multimodal time-series samples representing simulated children aged from 3–10 years. The dataset includes rich motion, acoustic, physiological, and contextual sensor streams across five annotated play behavior categories. The dataset simulates a naturalistic play environment with varying motion intensity, social proximity, and expressive characteristics. The data was split into 70% training and 30% testing. It consists of thorough, time-stamped recordings of children's vocalizations, body movements, and physiological conditions through different kinds of play. The information is arranged in synchronized records that show a child's vocalizations, body language, and emotional cues in a natural play environment. To facilitate research and learning in the fields of education, psychological development, and intelligent play circumstances, each entry is assigned a specific category based on play behavior. There are several key features, which are represented in Table 2.

Table 2: Significant features determined in the dataset

Features	Descriptions
Observations of Multimodal	It comprises motion, auditory, physiological, and physical data that illustrate different facets of play behavior.
Behavior Labels	Five categories of play behavior have been identified: Parallel play, cooperative play, playing alone, aggressive behavior, and inactive play.
Time-Series Information	Real-time observation is simulated by providing a distinct timestamp from 2024 to each record.
Signals of Emotion and Interaction	Body posture, verbal activity, social proximity, and emotional markers like heart rate are all represented by features.
Research-Focus	Established to support the comprehension of behavioral and interpersonal patterns in child development for educators, researchers, and developers.

3.2 Data preprocessing

The process of converting unprocessed data into a format that is more appropriate for modeling and evaluation is known as information processing. Obtained information requires being cleaned, transformed, and integrated to enhance its quality and facilitate the system's comprehension and processing. It fixes anomalies like missing data, inconsistencies, and noise to prepare the data for neural network algorithms. When examining children's play behavior, data preprocessing is essential. Academics and professionals acquire more information about the complicated dynamics of children's play behavior that is utilized for directing activities, learning techniques, and child development support. Two preprocessing techniques, such as Kalman filtering and z-score

normalization, are employed to evaluate the children's playing behavior.

3.2.1 Kalman Filtering to reduce noise

A Kalman filtering is an effective data preparation technique that smooth the noisy sensor streams, such as motion trajectories, object interactions, and ambient cues to ensure childrens play behavior. The Kalman filtering is appropriate for low-power edge devices used in classrooms or rehabilitation facilities due to its computational efficiency. The data instance's value is calculated by the Kalman filter using the observed value of the present instance and the known estimated value of the preceding moment. The Kalman filter is a probability distribution issue that determines the probability of the future by utilizing probability distribution and prior values. Using the state space technique, the Kalman filter characterizes the system's dynamic properties. The filter operates through two essential steps like prediction and correction. The state-space model in Equation (1).

$$A_{l+1} = XA_l + Y\mu_l + \omega_l \quad (1)$$

Where, ω_l is the process noise or disturbance, A is the actual system state, the state transition matrix is indicated as XA_l, and the control matrix is $Y\mu_l$, along with the control variable (μ).

Kalman filter allows the model to estimate child motion or posture even when sensors momentarily drop, fluctuate, or report inconsistent values. An error (f_{l+1}) calculation is indicated in Equation (2), and Equation (3) represents the uncertainty estimation ($Q_{l+1|l+1}$).

$$f_{l+1} = A_{l+1} - \hat{A}_{l+1|l+1} \quad (2)$$

$$Q_{l+1|l+1} = F(f_{l+1}f_{l+1}^S) = F\left(\left(A_{l+1} - \hat{A}_{l+1|l+1}\right)\left(A_{l+1} - \hat{A}_{l+1|l+1}\right)^S\right) \quad (3)$$

Exploring the position by employing the mathematical evaluation is presented in Equation (4).

$$\hat{A}_{l+1|l} = A_{l+1} - \widehat{XA}_{l|l} + Y\mu_l \quad (4)$$

The expected system state vector at time step l is represented by $\widehat{XA}_{l|l}$. The noise process uncertainty is denoted in Equation (5).

$$P_{l+1} = F(\omega_l + \omega_{l+1}^S) \quad (5)$$

The covariance matrix of the noise process is represented by P_{l+1}, whereas the noise process is represented by ω_{l+1}^S. Updated covariance is determined in Equation (6), and the uncertainty measure is calculate.

$$Q_{l+1|l+1} = (K - L_{l+1}G)Q_{l+1|l} \quad (6)$$

Where, covariance matrix is $Q_{l+1|l}$. It helps to refine the prediction by using real sensor measurements. Estimation of Kalman gain and the updated positions are indicated by Equations (7-8).

$$L_{l+1} = Q_{l+1|l}G^s\left(GQ_{l+1|l}G^s + Q_{l+1}\right)^{-1} \quad (7)$$

$$\hat{L}_{l+1|l+1} = \hat{L}_{l+1|l} + L_{l+1}\left(W_{l+1} - G\hat{A}_{l+1|l}\right) \quad (8)$$

Where, W_{l+1} is a measurement $\hat{A}_{l+1|l}$ is an anticipated system's state vector at time step l, and $\hat{L}_{l+1|l+1}$ is an assessed state vector of the system with time step $l+1$.

3.2.2 Z-score normalization to enhance consistency

The data preprocessing method, Z-score normalization, frequently referred to as standardization, converts data so that its mean is zero and its standard deviation is one. In this procedure, the
data is transformed into a unit variance and positioned at zero. Equation (9) denotes the z-score calculation.

$$Z - score = \frac{A-\mu}{\sigma} \quad (9)$$

Where, A represents the initial value, mean and standard deviation are represented by μ and σ.

3.3 Feature extraction

The process of turning incomplete information into a collection of new, important features that are more appropriate for predictive algorithms is known as feature extraction. It intends to enhance model performance, facilitate data representation, and lower dimensionality. Alternatively, selecting particular portions of the original features involves combining or altering existing characteristics to create new ones. The procedure of identifying and measuring specific features of a child's play that are subsequently utilized for evaluation, categorization, or other uses is known as feature extraction in terms of children's play behavior. Characteristics of the play activity, including play category, interactions with others, involvement level, and physical motions, are

represented by these features, which are easily obtained from a variety of data sources, including audio, video records, and sensor data. The Lightweight ResNet model is utilized in the research to extract the significant features from the children's playing behavior.

3.3.1 Lightweight ResNet model

Based on one of the existing training techniques, the ResNet50 model is employed in the playing behavior of children assessment. The collection of a pre-trained model is used to establish a model that fails to comprehend anything about images. By allowing training with fewer data sets, ResNet50 reduces the computing expenses. The ResNet50 model's input layer is configured to receive 224,224,1 values from the data set. After the input layer, the convolution layer's values are then updated. Figure 2 shows the entire architecture of ResNet50, collectively with the additional levels.

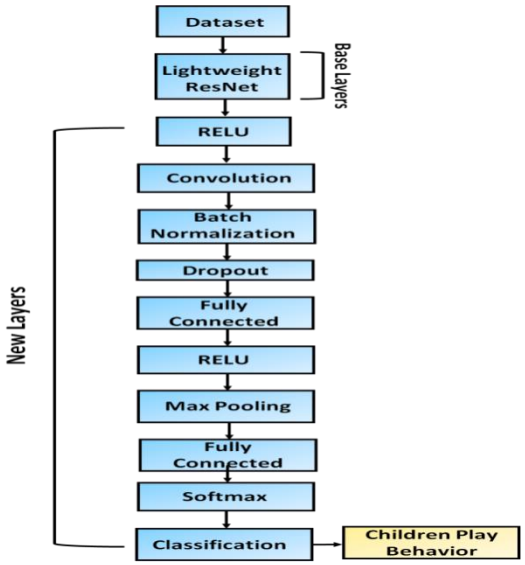


Figure 2: Entire Architecture of ResNet50

The new model is built from the ResNet50 model's input, convolution, activation, pool, fully-connected, softmax, and classification layers. Two new fully connected layers are generated; batch normalization is implemented for input values as well as stability and speed are improved. The output layer's fully connected structure uses Softmax activation for data classification. Dropout prevents the model from remembering training data.

Input Layer: This layer serves as the model's primary layer, and this layer's highest selection of input image sizes increased the amount of storage needed while extending the training and testing durations. As a result, all architectures of the input layer are determined to be $224 * 224 * 1$.

Activation Function: The activation layer is another designation for ReLU. Negative values in the input data are assigned to zero in the outcome. The network operates more quickly when its negative dimension value is zero. This investigation made use of the ReLU activation function. In Equation (10), the ReLU activation function is provided.

$$E(a) = \begin{Bmatrix} 0, a < 0 \\ a, a \geq 0 \end{Bmatrix} \tag{10}$$

By enabling its lower computational demands than other functions, the ReLU layer is more supported.

Layer of Convolution: The foundation of CNN networks is the convolution layer and also known as the transformation layer. Convolution is the procedure of applying filters to all layers. This layer's specified filters have $N \times N$ sizes. Equation (11) provides the convolution that consist of linear filters.

$$(g_l)_{ji} = (Z_l * a)_{ji} + y_{ji} \tag{11}$$

Where a represents the input data, (j, i) represents the pixel point index, l represents index of the feature map, Z and y represents weighing parameters, and $(g_l)_{ji}$ represents the feature map's output value.

Normalization: The network's efficiency is increased by the normalization procedure. The data on additional layers may have different dimensions. According to Equations (12) and (13), the normalization process is as follows:

$$a^l = \frac{a^l - F(a^l)}{\sqrt{Var(a^l) + \varepsilon}} \tag{12}$$

$$b^{(l)} = \gamma^l a^l + \beta^l \tag{13}$$

Where $b^{(l)}$ denotes the input's dimension and $F(a^l)$ denotes the dimension's average. The definition of the standard deviation is $\sqrt{Var(a^l)+\varepsilon}$. There are two learnable variables, γ and β.

Dropout Layer: A lot of data is used in deep learning to train networks. Therefore, the network has been trained when the memorization event is possible. It is necessary to remove certain nodes that stop the network from memorizing. Implement dropout to enhance network performance.

Fully-connected Layer: This layer is dependent on every field of the preceding layer. The Fully Connected Layer transforms the information from the previous layer into a one-dimensional matrix structure. There are possible variations in the variety of entirely interconnected layers that the architecture utilizes.

Pooling Layer: The input data size reduction and the computational complexity reduction are the primary objectives presented in this layer. The $N \times N$ size filters are selected in the pooling layer. The size of the completed image is determined by pooling, as demonstrated in Equations (14-16).

$$T = z2 * g2 * c2 \tag{14}$$

$$z2 = \frac{(z1-e)}{X+1} \tag{15}$$

$$g2 = \frac{g1-e}{X+1} \tag{16}$$

Where $z1$ represents the input width, $g2$ is the height, $c1$ indicates the image depth, e denotes the dimension of the filter, X determines the step counts, and the size of the data is indicated as T. In the suggested architecture, the pooling layer is maximum pooling.

Softmax Layer: In the classification process, it generates the probabilistic value using the previous layer's output. According to Equation (17), it calculates the values for every class. These possibilities estimate the classes using values ranging from 0 to 1.

$$Q(b = i|a; Z, y) = \frac{Exp^{A^{SZ_i}}}{\sum_{i=1}^{N} Exp^{A^{SZ_i}}} \tag{17}$$

Where the main class Z and y is a vector of weights. These processes make use of cross-entropy. Equation (18) provides the cross-entropy function that is most frequently used.

$$CrossEntropy = -\sum_a Q'(a) LogQ(A) \tag{18}$$

Whereas Q represents the actual production, Q' represents the expected output. Finally, images are categorized in the classification layer.

3.4 Assessing the children's play behavior through the autoencoder-based long short-term memory (AE-LSTM)

The integration of AE and LSTM model is used for analyzing and understanding children's play behavior. The AE-LSTM models help to adjust the children's specific variations by identifying distinctive patterns and behavioral abnormalities, which is crucial for customized monitoring and evaluation of children's. The AE-LSTM system on edge devices further enhances its practical utility of computational workload, performed close to the data source for reducing latency, and enabling real-time feedback. It makes the ability of the research to identify patterns in time-series data, such as the play behavior of children, effective and in real-time. In the AE-LSTM model, both encoder and decoder weights were jointly optimized during sequence learning. Two losses were trained together: (1) the autoencoder's reconstruction loss for learning compact temporal representations, and (2) the Softmax cross-entropy loss for play-behavior classification. Edge computing ensures to be local, for eliminating dependence on cloud connectivity and reducing response time for real-time feedback. Lightweight ResNet effectively extracts multimodal features while maintaining computational efficiency. The AE-LSTM model enhances sequential behavior analysis by combining dimensionality reduction with robust temporal modeling.

3.4.1 Long short-term memory (LSTM)

The Recurrent Neural Networks (RNNs) of the LSTM type are made to recognize and remember long-term dependencies in sequential input. It manages information flow through storage cells and gating mechanisms that make it useful for behavior evaluation tasks that involve the modeling of time-series data like voice, movement, or physiological signals from the children during playtime. Time series data is interpreted using a particular kind of computer-based RNN architecture called LSTM. RNNs have difficulty with gradient difficulties and long-term dependencies, which affects the capacity to accurately analyze complicated and sequential data on children's play activity. By utilizing gated memory cells, LSTM addressed the RNN's gradient difficulties and made it possible to accurately represent long-term dependencies. The infrastructure is more capable of detecting irregularities in assessing the children's behavior during play when the LSTM architecture is combined with an AE that assigns significance to important sequences.

As compared to conventional transfer networks, LSTM's feedback interactions between hidden components linked to certain time steps allow for the development of long-term sequence dependency and the

forecasting of interaction labels determined by the sequence of previous activities. To address the issue of diminishing and exploding gradients that occur during the training of conventional RNNs, LSTMs were developed. Updates are made to the data stored in the memory cell of the LSTM unit through the input, forget, and output gates. During random intervals, the factor maintains values, and the three gates control the flow of information from and to the factor. The LSTM's single unit is represented in Figure 3.

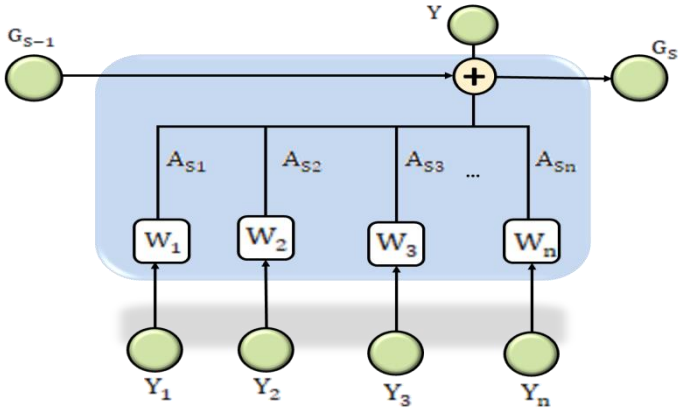


Figure 3: Design of LSTM Unit

Following Equations (19-24) is an approach to computing each cell in an LSTM.

$$a = \begin{bmatrix} G_{S-1} \\ A_S \end{bmatrix} \tag{19}$$

$$G_S = \sigma(W_G \cdot a + Y_G) \tag{20}$$

$$J_S = \sigma(W_J \cdot a + Y_i) \tag{21}$$

$$R_S = \sigma(W_R \cdot a + Y_R) \tag{22}$$

$$D_S = G_S \odot D_{S-1} + J_S \odot \tanH(W_D \cdot a + Y_D) \tag{23}$$

$$G_S = R_S \odot \tanH(D_S) \tag{24}$$

Where $W_J, W_G, W_R \in \mathbb{R}^{D \times 2D}$ are positioned in training, the weighted measures and Y_J, Y_E, $Y_R \in \mathbb{R}^D$ biases of the LSTM are learned, comprising three gates' transformations. Variable σ is the sigmoid function and element-wise multiplication is represented by $\odot$. The LSTM cell unit's inputs are contained in the vectors A_S. The vector of the hidden layer is Z_S. After linearizing the sentence into a vector with a size equal to the number of class labels, insert the final hidden vector n to indicate the phrase as a Softmax layer. Class labels that are neutral, negative, and positive are utilized.

3.4.2 Autoencoder (AE)

To improve proactive children's play behavior in real-time environments, the AE is used to obtain compact and resilient representations of time-series play behaviors. An AE is ideal for this task. It encodes the input sequences into a compact latent space while removing inconsequential information, and then recovers the original input data while reducing reconstruction loss via the decoding structure. The AE's efficient representation allows it to catch hidden patterns, which can be useful in spotting performance abnormalities. The technique is divided into three stages: encoding the input into a compressed latent space, decoding it to rebuild the input, and decreasing reconstruction loss.

Encoding: To encode high-dimensional input for assessing the children's play behavior in real-world environments, the encoder maps the input vector $x \in R^m$ into a compressed latent representation h. It is achieved using Equation (25).

$$h = f_1(w_i x + b_i) \tag{25}$$

The encoder weights and biases are denoted by wi and b_i, respectively, while the activation function is represented by f_1. This stage removes inconsequential differences while preserving fundamental structural elements.

Decoding: The decoder remaps the compressed representation into a reconstructed input $\hat{x}$ to recognize children's play behavior. This transformation is represented as follows in Equation (26).

$$\hat{x} = f_2(w_j h + b_j) \tag{26}$$

Where w_j and b_j are the weights of decoders and bias, and the activation function (f_2) is applied to reconstruct the input structure.

Reconstruction Loss: In quantifying reconstruction deviations for assessing children's play behavior in real-world conditions, the model computes loss (L) between inputs x and its reconstruction $\hat{x}$ as indicated by Equation (27).

$$L(x - \hat{x}) = \frac{1}{n}\sum_{n=1}^{n} |\hat{x}_t - x_t| \tag{27}$$

Where, x is the actual input data disregarding the observed system behavior $\hat{x}$ is the reconstructed output from the autoencoder, and n is the quantity of training illustrations. The loss function supports quantifying differences between actual and predicted behavior, enabling the detection of children's play behaviors relatively earlier than otherwise observable. This is refined by a position-aware formulation, Equation (28).

$$x_i = \frac{1}{n}\sum_{n=1}^{n} |\hat{x}_i - x_i|, \ \ where\ n = \begin{cases} N & if\ i \le \frac{N+1}{2} \\ n - i + 1 & if\ i > \frac{N+1}{2} \end{cases} \tag{28}$$

Variable x_i represents reconstruction error for the i model, and $\hat{x}_i$ is the predicted rate, n is the total sequence length, and N is adjusted per contextual importance over time. This weighted-based mechanism is validated by providing attention to children's play behavior in real-world circumstances. Overall reconstruction loss across time series is presented by Equation (29).

$$loss = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{29}$$

Where, x_i is the reconstruction loss and N is the full sequence length. Algorithm 1 shows the AE-LSTM algorithm.

Algorithm 1: AE-LSTM

```
Input:
  D = {Training, Validation, Test} datasets
  Model Parameters:
    Encoder: wi, bi
    Decoder: wj, bj
    LSTM gates:
      W_G, W_J, W_R, W_D
      Y_G, Y_J, Y_R, Y_D
  Hyperparameters:
    lr, epochs, batch_size, clip_norm, patience
Initialize:
  optimizer ← Adam({all parameters}, lr)
  best_val_loss ← ∞
  no_improve ← 0
Function Encode(x):
  For t = 1 ... N:
    h_t ← f1(wi * x[t] + bi)
  return {h_1 ... h_N}
Function LSTM_Forward(H):
  Z_0 ← 0 ; D_0 ← 0
  For S = 1 ... N:
    a ← concat(Z_(S − 1), H[S])
    G_S ← sigmoid(W_G * a + Y_G)
    J_S ← sigmoid(W_J * a + Y_J)
    R_S ← sigmoid(W_R * a + Y_R)
    D_S ← (G_S ⊙ D_(S − 1)) + (J_S ⊙ tanh(W_D * a + Y_D))
    Z_S ← R_S ⊙ tanh(D_S)
  return Z_S, {Z_1 ... Z_N}, {D_1 ... D_N}
Function Decode(H):
  For t = 1 ... N:
    x_hat_t ← f2(wj * H[t] + bj)
  return {x_hat_1 ... x_hat_N}
Function Position_Aware_Loss(x, x_hat):
  N ← length(x)
  loss_sum ← 0
  For i = 1 ... N:
    if i ≤ (N + 1)/2:
      n_i ← N
    else:
```

```
      n_i ← N − i + 1
    e_i ← mean(|x_hat[i] − x[i]|)
    x_i ← (1 / n_i) ∗ e_i
    loss_sum ← loss_sum + x_i
  return loss_sum / N
Training Loop:
For epoch = 1 ... epochs:
  Shuffle Training data
  For each batch B:
    X_batch ← inputs in B
    For each sequence x in X_batch:
      H ← Encode(x)
      Z_final, Z_seq, D_seq ← LSTM_Forward(H)
      x_hat ← Decode(H)
      L_seq ← Position_Aware_Loss(x, x_hat)
      Accumulate L_seq
    batch_loss ← mean(L_seq)
    optimizer.zero_grad()
    Backprop(batch_loss)
    Clip_Gradients(all_parameters, clip_norm)
    optimizer.step()
  val_loss ← Evaluate(D.Validation)
  If val_loss < best_val_loss:
    best_val_loss ← val_loss
    Save_Checkpoint(model_parameters)
    no_improve ← 0
  Else:
    no_improve ← no_improve + 1
    If no_improve = patience:
      Break
Evaluation (Test Phase):
Load_Checkpoint(best_model)
For each sequence x in Test set:
  Compute Encode → LSTM_Forward → Decode
  Compute Position_Aware_Loss
Return final test_loss and metrics
Final Evaluation:
  Load best checkpoint
  test_loss, test_metrics = Evaluate(D_test, model_parameters)
  RETURN best_model, test_loss, test_metrics
Procedure Evaluate(D_split, params):
  Set model to eval mode (disable dropout, etc.)
  losses = []
  pred_labels = [], true_labels = []
  FOR each sequence x (and optionally y) in D_split:
    compute x_hat_seq and final hidden Z_final (no gradient)
    compute per − sequence position − aware loss x_i as in training
    seq_loss = mean_over_i(x_i)
    append seq_loss to losses
    IF labels present:
      logits = Linear(Z_final)
      y_pred_label = argmax(softmax(logits))
      append y_pred_label to pred_labels
      append y to true_labels
  avg_loss = mean(losses)
  IF labels present:
    metrics = compute_metrics(true_labels, pred_labels)
  ELSE:
    metrics = {}
  RETURN avg_loss, metrics
```

Learning compact, sequential representations of temporal data while capturing long-range dependencies is accomplished with AE-LSTM. This research allows for real-time monitoring of children's play behavior by encoding multimodal sensor inputs and decoding patterns to accurately identify social interactions, activity categories, and emotional states in dynamic and noisy conditions. Several hyperparameters utilized in the research are explored in Table 3.

Table 3: Hyperparameters and Values for AE-LSTM model configuration

Training Epochs	Learning Rate	Batch Size	Latent Dimension (AE)	LSTM Hidden Units	Dropout Rate	Sequence Length
10 Epochs	0.001	32	32	64	0.1	30
20 Epochs	0.001	32	48	96	0.15	40
30 Epochs	0.001	64	64	128	0.20	50
40 Epochs	0.0008	64	64	128	0.25	60
50 Epochs	0.0005	64	64	128	0.30	60

4 Experimental results

The research intended to provide a real-time assessment of children's play behavior. The following phases provide a detailed explanation of the research results. The proposed platform is designed to run on lightweight edge devices, which minimizes hardware cost and eliminates the need for high-performance servers. Sensors such as RGB-D cameras, wearable units, and microphones are selected based on low-power, commercially available modules to ensure affordability for classrooms and homes.

4.1 Evaluation criteria

This evaluation criteria section provides the evaluation outcomes of the proposed AE-LSTM method in various feature parameters like hesart rate distribution through play behavior, skin temperature during activity levels, comparison of heart rate and activity levels, evaluation of activity level and proximity to peers, pitch mean and standard deviation distributions, and physiological signals distribution. The discussion below discusses the evaluation features of the proposed AE-LSTM approach.

Figure 4 represents the engagement and feedback relevance of children play behavior. It indicates the median and variable heart rates, providing information about how the body responds to various forms of play. The AE-LSTM model uses temporal patterns in multimodal data including heart rate, posture, speech tone, and motion to classify and interpret play behavior dynamically. It helps to physiological variation across spontaneous play states. By using the information to track mental and physical states, the AI platform provides real-time feedback while playing and correlates internal signals with behavior.

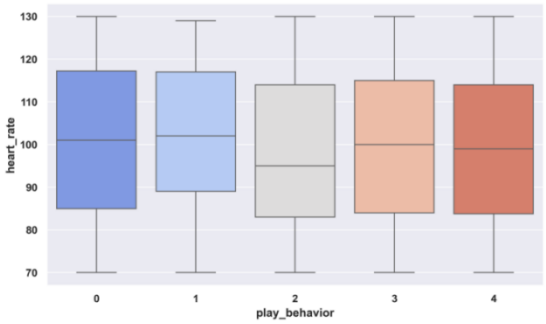


Figure 4: Engagement and feedback relevance of children play behavior

Depending on the type of play, heart rates might vary from 85 to 145 beats per minute. Running and other high-intensity activities exhibit median heart rates of about 130 bpm, whereas peaceful play activities have median heart rates of about 90 bpm. The inference of tension or excitement in real time is influenced by the values. The AI model's physiological-behavioral mapping is improved by significant interquartile ranges, which show behavioral variability.

Skin Temperature during Activity Levels: Figure 5 shows how children's skin temperatures change with different degrees of activity. Increased physical effort or emotional stimulation is frequently indicated by elevated temperatures. The AE-LSTM model model leverages temporal patterns in multimodal data including skin temperature to classify and interpret play behavior in real time. To determine emotional states, wearable sensors capture these physiological signals and incorporate them into the platform. This allows for real-time behavior classification and adaptive feedback.

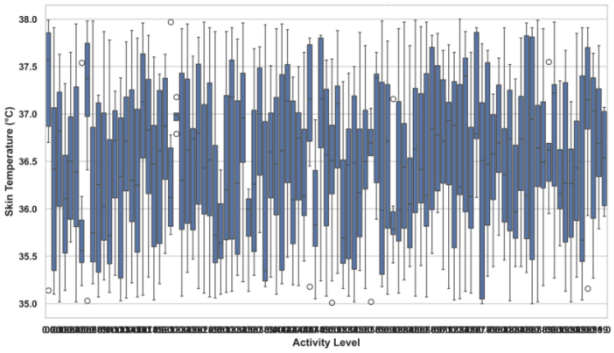


Figure 5: Evaluation of skin temperature through activity levels

Skin temperatures range from 32.0 to 36.5°C. The median temperature is near 36.0°C for high activity levels and closer to 33.5°C for passive behavior. Emotional changes or exertion are reflected in fluctuations. The adaptive feedback loop for children's comfort and stress detection is supported by these physiological readings, along with motion information for contextual behavior recognition.

Figure 6 depicts the bubble plot of physiological-behavioral mapping, which is categorized by color according to various play behaviors. The AE-LSTM model used to enhance the subtle temporal patterns that differentiate between levels of spontaneous play. It shows how multimodal physiological and behavioral information is used to enhance the cognitive performance.

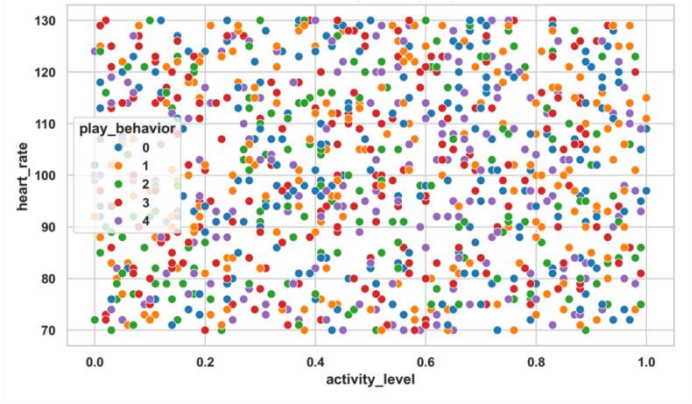


Figure 6: Bubble plot of physiological-behavioral mapping

There is a positive correlation between heart rate (85–145 bpm) and activity level (range: 0–10). Claiming and other play behaviors show top-right clustering (activity > 8, heart rate > 130 bpm), whereas quiet activities are located close to the origin. Physiological-kinematic synchronization is improved by the connection in the classification of multimodal behavior.

Activity Level and Proximity Peers: The connection between children's levels of activity and the way they are connected to their classmates is demonstrated in Figure 7. The AE-LSTM model learns temporal and spatial patterns to infer behavioral states in real time. By showing how the social and physical aspects of play co-vary, expose behavioral clusters that could represent cooperative, solitary, or transitional play styles. It allows the AI to identify social involvement levels in real time.

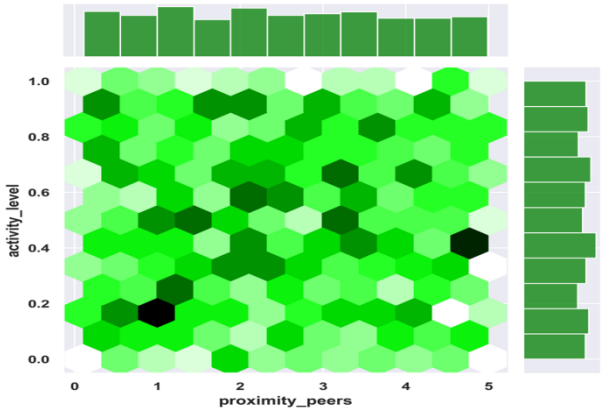


Figure 7: Results of activity levels and proximity peers

The range for proximity is 0.5–3.0 meters, while the range for activity level is 0–10. Activity > 7 and high densities at <1.5 m proximity indicate active, social play. Solitary or passive intervals are highlighted by sparse areas over long distances and low activity levels. Based on the outcomes, the platform categorizes different kinds of interactions, such as self-sustaining and collaborative.

Physiological Signals Distribution across play behavior types was illustrated in Figure 8.The AE-LSTM network leverages inputs to model temporal dependencies and reconstruct latent behavioral patterns. By capturing both individual and joint distributions of physiological metrics, the system enhances its ability to infer emotional and physical engagement levels. Significant differences in physiological indicators associated with various play behaviors are revealed by the diagonal plots, which show the kernel density calculation of all characteristics by behavior type.

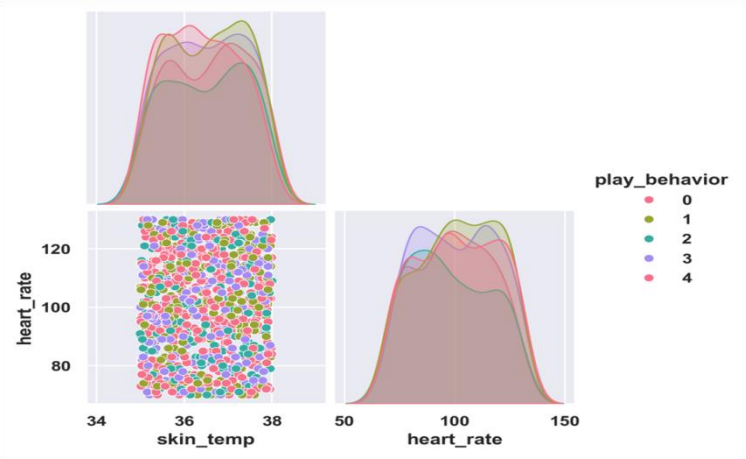


Figure 8: Physiological signals distribution across play behavior types

Five play behavior classes' relationships between skin temperature (34–38°C) and heart rate (60–150 bpm) are presented. There is a unique physiological characteristic for every behavior determined in the results. Behavior 0 is linked to higher heart rates, although behavior 2 leads to lower temperatures.

Pitch Mean and Standard Deviation Distributions: Pitch mean and pitch standard deviation distributions from children's speech data are displayed in Figure 9. By capturing both central tendency and variability in pitch, the system can detect shifts in emotional tone and engagement. This visualization shows how speech-based metrics enrich the system's ability to interpret and support child development through intelligent, responsive play environments. These audio features support real-time AI analysis by assisting in the inference of behavioral and emotional cues during play.

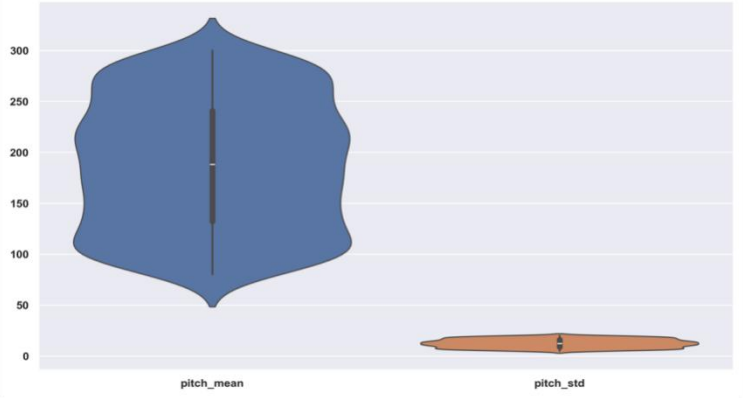


Figure 9: Determination of mean and standard deviation distributions

The voice pitch varies during play, as indicated by the pitch mean, which varies from roughly 75 Hz to 310 Hz. The pitch standard deviation (pitch_std) shows constant patterns of vocal variability, remaining densely emphasized between 5 Hz and 15 Hz.

Table 4 presents the performance of different models' confidence intervals. The confidence intervals further confirm the statistical reliability of these results, highlighting AE-LSTM's superior consistency and effectiveness. AE-LSTM outperforms all other models, achieving the highest scores in all metrics, indicating robust and reliable predictions.

Table 4: Evaluation metrics of AE-LSTM and baseline models confidence intervals

Model	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)
AE-LSTM	0.968 – 0.982	0.960 – 0.976	0.956 – 0.972	0.950 – 0.968
LSTM	0.833 – 0.865	0.819 – 0.851	0.805 – 0.839	0.811 – 0.845
GRU	0.816 – 0.848	0.800 – 0.834	0.788 – 0.822	0.794 – 0.828
BiLSTM+Attention	0.857 – 0.887	0.844 – 0.874	0.835 – 0.867	0.840 – 0.870

Table 5 presents the results of significance testing for the AE-LSTM model across using a significance level. These results confirm that the observed performance of the AE-LSTM model demonstrates robust and reliable predictive capability across all evaluated metrics.

Table 5: Statistical significance analysis of AE-LSTM performance metrics

Metric	AE-LSTM Value ($\hat{p}$)	z-Statistic	p-Value	Significance ($\alpha = 0.05$)
Accuracy	0.975	42.31	$p < 0.0001$	Significant
Precision	0.968	41.28	$p < 0.0001$	Significant
Recall	0.964	40.74	$p < 0.0001$	Significant
F1-Score	0.959	40.07	$p < 0.0001$	Significant

4.2 Comparison phases

The research compares the proposed AE-LSTM method with various existing techniques, such as Gated Recurrent Unit (GRU) [22], Bidirectional LSTM (BiLSTM) [22], and BiLSTM+Attention [22], to assess the playing behavior. Table 6 determines the comparison evaluation of proposed and existing methods with F1-score, precision, recall and accuracy. The performance matrix's formula and definitions are provided in Table 7.

Table 6: Formulas and definitions of performance matrices

Metrics	Definitions	Equations
Accuracy	The proportion of accurate true positive and true negative forecasts overall.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	It is a proportion of real positive forecasting over all the positive predictions.	$\frac{TP}{TP+FP}$
Recall	The percentage of real positive cases that a prediction model accurately classifies as positive is known as recall.	$\frac{TP}{TP+FN}$
F1 Score	It is the harmonic mean of precision and recall. It measures the balance between both metrics.	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

Table 7: Comparison of outcomes of AE-LSTM and exiting methods

Models	$Accuracy$	$Precision$	$Recall$	$F1-score$
LSTM [22]	0.849	0.835	0.822	0.828
GRU [22]	0.832	0.817	0.805	0.811
BiLSTM+Attention [22]	0.872	0.859	0.851	0.855
AE-LSTM [Proposed]	0.975	0.968	0.964	0.959

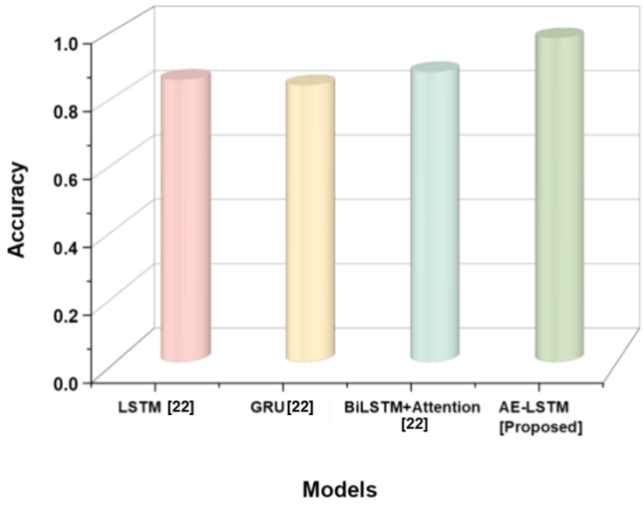


Figure 10: Outcomes of the accuracy metric in play behavior

Figure 10 depicts the outcomes determined by the accuracy. The proposed AE-LSTM approach provides an accuracy of 0.975, whereas GRU shows 0.832 accuracy, BiLSTM+Attention has 0.872, and LSTM provides 0.849 accuracy. Based on the results, the proposed AE-LSTM method has a high accuracy to compact latent representations from multimodal play behavior data and to effectively capture physiological signal distributions.

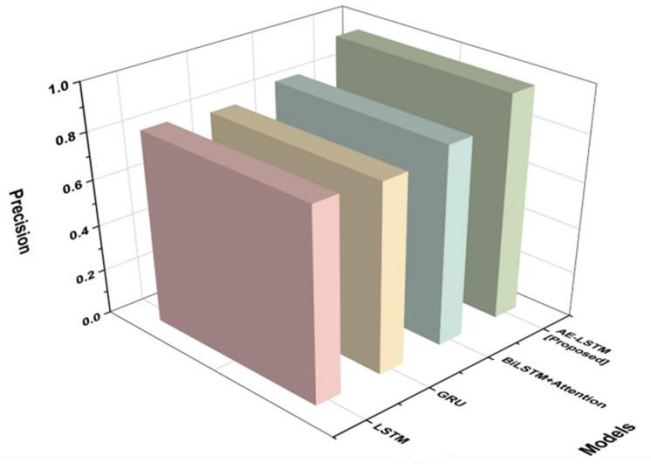


Figure 11: Result of play behavior with precision

Figure 11 shows precision findings. Existing methods provide precision outcomes (LSTM has 0.835, BiLSTM+Attention has 0.859, and GRU has 0.817). The precision of the AE-LSTM method is 0.968, and it indicates that the AE-LSTM technique has more efficiency, and reliability in children's play behavior than other existing models.

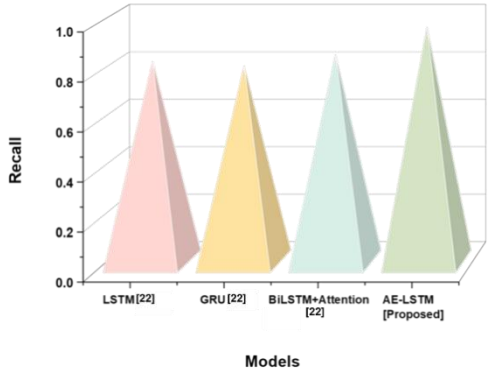


Figure 12: Visual depiction of recall results

Figure 12 demonstrates the results of recall. The proposed AE-LSTM approach indicates essential results in terms of recall (0.964), indicating its strong capability in accurately identifying children's play behavior pattern. Outcomes of the research demonstrate that the proposed AE-LSTM approach is more significant with a recall result the existing techniques like GRU (0.805), LSTM (0.822), and BiLSTM+Attention (0.851).

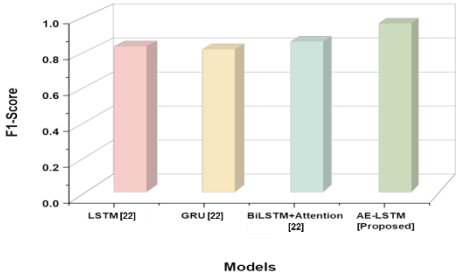


Figure 13: Evaluation outcomes of F1-score

Estimation of the F1-score is displayed in Figure 13. The AE-LSTM method has a 0.959 F1-score, BiLSTM+Attention has 0.855, GRU has 0.811, and LSTM has a 0.828 F1-score. The proposed AE-LSTM method explores superior results compared to the traditional approaches, as determined by the research findings. By effectively modeling temporal patterns and physiological signal distributions, the AE-LSTM captures subtle

behavioral variations, enabling more accurate classification.

Assessing the behavior of playing children was the focus of the research. The evaluation of the research compared the proposed method with existing techniques, and certain limitations were observed in the existing methods in assessing the play behaviors. The LSTM [22] model might have trouble in recognizing long-range temporal connections. The computing capacity of edge devices could limit the accuracy of real-time inference and model complexity. The computational depth of GRU [22] was insufficient to capture complex temporal connections. Although it has superiority in contextual training, BiLSTM [22] was less appropriate for real-time applications due to its increased computing complexity and latency. The BiLSTM+Attention [22] raises attention on vital features, whereas it was expensive for edge distribution and issues from over-fitting with small training data. These models have a general difficulty with dynamic and natural environments in children's play. Our system balances latency, memory usage, and real-time processing by deploying lightweight ResNet feature extraction and the AE-LSTM architecture directly on edge hardware. Unlike GRU, LSTM, or BiLSTM-Attention models, the AE-LSTM compresses multimodal inputs into compact latent representations, reducing memory load while maintaining strong temporal modeling. Because computation occurs locally, inference latency is significantly lower than cloud-based models, enabling immediate feedback. Existing observation methods for children's play are largely manual, subjective, and unsuitable for real-time interpretation. To address these shortcomings, the proposed AE-LSTM model enhances the real-time understanding of children's natural play. The AE model helps to compress and denoise multimodal features, whereas LSTM captures temporal patterns and emotional cues. Overall, the conceptualized model AE-LSTM will facilitate a real-time analysis of complex sensory play behavior, which provides robust temporal modeling.

5 Conclusions

1.The purpose of the research was to create and implement a multimodal sensor and feedback platform by utilizing edge computing and real-time AI to monitor, evaluate, and assist children's play behavior. The multimodal play behavior dataset with various sensors was obtained for the performance. Kalman filtering and normalizing techniques were used to pre-process data to increase consistency and minimize noise. Real-time behavior analysis was performed with an AE-LSTM network, while feature extraction was accomplished with a Lightweight ResNet model. To provide low-latency processing, local data storage, and privacy protection, the entire system has been deployed on edge devices. To increase engagement, the device provided real-time input via physical and visual signals. According to the experimental results, play behavior classification, emotional state detection, and peer interaction identification were all achieved with highly accurate results. Comparison of the proposed method demonstrates significant results in terms of accuracy (0.975), precision (0.968), recall (0.964), and F1-score (0.959). A real-time interpretation of children's play behavior, the proposed model supports educators in understanding engagement levels, social interaction patterns, and emotional cues during learning activities. Therapists gain continuous behavioral monitoring that supports early detection of developmental needs. Researchers benefit from reliable, unobtrusive multimodal analytics that capture natural play behavior accurately.

5.1 Limitations and future scopes

Children's behavior variability, sensor position limitations, and the requirement for frequent validation in dynamic playing environments limit the system's efficacy. The scalability of the system across different cultural or contextual settings remains challenge, which could affect its generalizability. The long-term deployment feasibility, include issues like battery life, device comfort, and overall cost, which are essential for practical implementation. Additionally, it collects sensitive data from children, ethical and privacy considerations lead to potential risks. Adaptability in a variety of play environments, emotional recognition skills, and the incorporation of adaptive learning models will represent the main areas of future research. To assess the platform's scalability and developmental impact in larger educational or clinical contexts, further long-term investigations are needed.

Data availability

The datasets generated and/or analysed during the current study are available in the Kaggle repository, [https://www.kaggle.com/datasets/ziya07/multimodal-play-behavior-dataset/data].

Ethics approval

The data analyzed in this study originate from the publicly available "Multimodal Play Behavior Dataset" on Kaggle. Therefore, no new ethics committee approval or informed consent was required for our analysis.

Consent to participate

Not applicable, as the study is entirely based on publicly available secondary data with no identifiable personal

Funding

This work was supported by 2024 construction project for the "1112" teaching project of Shaanxi College of Apparel Engineering, "Game Design for Preschool Children Offline Specialized Course" (No. 2024TSKC020)

References

[1] Hyun, S., McWayne, C.M. and Smith, J.M., 2021. “I see why they play”: Chinese immigrant parents and their beliefs about young children's play. Early Childhood Research Quarterly, 56, pp.272-280. https://doi.org/10.1016/j.ecresq.2021.03.014

[2] Passmore, A.H. and Hughes, M.T., 2021. Exploration of play behaviors in an inclusive preschool setting. Early Childhood Education Journal, 49(6), pp.1155-1164. https://doi.org/10.1007/s10643-020-01122-9

[3] Li, S., Sun, J. and Dong, J., 2022. Family socio-economic status and children’s play behaviors: The mediating role of home environment. Children, 9(9), p.1385. https://doi.org/10.3390/children9091385

[4] Wu, S., Pan, C., Yao, L. and Wu, X., 2022. The impact of the urban built environment on the play behavior of children with ASD. International journal of environmental research and public health, 19(22), p.14752. https://doi.org/10.3390/ijerph192214752

[5] Dodd, H.F., Nesbit, R.J. and Maratchi, L.R., 2021. Development and evaluation of a new measure of children’s play: The Children’s Play Scale (CPS). BMC Public Health, 21(1), p.878. https://doi.org/10.1186/s12889-021-10812-x

[6] van Dijk-Wesselius, J.E., Maas, J., van Vugt, M. and van den Berg, A.E., 2022. A comparison of children's play and non-play behavior before and after schoolyard greening monitored by video observations. Journal of Environmental Psychology, 80, p.101760. https://doi.org/10.1016/j.jenvp.2022.101760

[7] Lennon, M., Pila, S., Flynn, R. and Wartella, E.A., 2022. Young children's social and independent behavior during play with a coding app: Digital game features matter in a 1: 1 child to tablet setting. Computers & Education, 190, p.104608. https://doi.org/10.1016/j.compedu.2022.104608

[8] Dankiw, K.A., Kumar, S., Baldock, K.L. and Tsiros, M.D., 2024. Do children play differently in nature play compared to manufactured play spaces? A quantitative descriptive study. International Journal of Early Childhood, 56(3), pp.535-554. https://doi.org/10.1007/s13158-023-00384-9

[9] Yang, J.T., Chen, C.I. and Zheng, M.C., 2023. Elevating children’s play experience: a design intervention to enhance children’s social interaction in park playgrounds. Sustainability, 15(8), p.6971. https://doi.org/10.3390/su15086971

[10] Cakan, A. and Acer, D., 2024. Analysis of preschool children’s outdoor play behaviors. Journal of Outdoor and Environmental Education, pp.1-27. https://doi.org/10.1007/s42322-024-00174-4

[11] López-Toribio, M., Hidalgo, L., Litt, J.S., Daher, C., Nieuwenhuijsen, M., Márquez, S., Berrón, A., Franch, B., García, B. and Ubalde-López, M., 2025. SOOPEN: design and assessment of a tailored systematic observation tool to evaluate outdoor play behavior among schoolchildren groups. Cities & Health, pp.1-12. https://doi.org/10.1080/23748834.2024.2439645

[12] Wilhelmsen, T., Lekhal, R., Rydland, V. and Coplan, R.J., 2025. Exploring the role of early childhood educators’ emotion socialization strategies in the development of young children’s social and non-social play behaviors. Early Childhood Research Quarterly, 73, pp.92-100. https://doi.org/10.1016/j.ecresq.2025.06.005

[13] Yesil, R., Erdiller Yatmaz, Z. and Metindogan, A., 2025. Exploring Children’s Play Culture and Game Construction: Role of Sociodramatic Play in Supporting Agency. Early Childhood Education Journal, 53(3), pp.703-716. https://doi.org/10.1007/s10643-023-01621-5

[14] Samuelsson, R., 2023. A shape of play to come: Exploring children's play and imaginaries with robots and AI. Computers and Education: Artificial Intelligence, 5, p.100173. https://doi.org/10.1016/j.caeai.2023.100173

[15] Ocak, C., Kopcha, T.J. and Dey, R., 2023. An AI-enhanced pattern recognition approach to temporal and spatial analysis of children's embodied interactions. Computers and Education: Artificial Intelligence, 5, p.100146. https://doi.org/10.1016/j.caeai.2023.100146

[16] Nopembri, S., Mulyawan, R., Fauziah, P.Y., Kusumawardani, E., Susilowati, I.H., Fauzi, L., Cahyati, W.H., Rahayu, T., Chua, T.B.K. and Chia, M.Y.H., 2023. Time to play in Javanese preschool children—An examination of screen time and playtime before and during the COVID-19 pandemic. International Journal of Environmental Research and Public Health, 20(3), p.1659. https://doi.org/10.3390/ijerph20031659

[17] de Castro Rodrigues, D., de Siqueira, V.S., da Costa, R.M. and Barbosa, R.M., 2022. Artificial Intelligence applied to smart interfaces for children’s educational games. Displays, 74, p.102217. https://doi.org/10.1016/j.displa.2022.102217

[18] Shawcroft, J.E., Gale, M., Workman, K., Leiter, V., Jorgensen-Wells, M. and Jensen, A.C., 2022. Screen-play: An observational study of the effect of screen media on Children's play in a museum setting. Computers in Human Behavior, 132, p.107254. https://doi.org/10.1016/j.chb.2022.107254

[19] Faraz, A., Mounsef, J., Raza, A. and Willis, S., 2022. Child safety and protection in the online gaming ecosystem. Ieee Access, 10, pp.115895-115913. https://doi.org/10.1109/ACCESS.2022.3218415

[20] Gashaj, V., Dapp, L.C., Trninic, D. and Roebers, C.M., 2021. The effect of video games, exergames and board games on executive functions in kindergarten and 2nd grade: An explorative longitudinal study. Trends in Neuroscience and Education, 25, p.100162. https://doi.org/10.1016/j.tine.2021.100162

[21] Foweather, L., Crotti, M., Foulkes, J.D., O'Dwyer, M.V., Utesch, T., Knowles, Z.R., Fairclough, S.J., Ridgers, N.D. and Stratton, G., 2021. Foundational movement skills and play behaviors during recess among preschool children: A compositional analysis. Children, 8(7), p.543. https://doi.org/10.3390/children8070543

[22] Liang, R., Ye, Z., Liang, Y. and Li, S., 2025. Deep Learning-Based Play Behavior Modeling and Game Interaction System Optimization Research. https://doi.org/10.20944/preprints202505.2198.v1

Appendix

Symbol	Explanation / Meaning
A_l, A_{l+1}	Actual system state at time step l of $l+1$.
X	State transition matrix governing state dynamics.
Y	Control matrix.
μ_l	Control input at time step l .
ω_l	Process noise or disturbance at time step l.
f_{l+1}	Error between actual and estimated state at time $l+1$.
$(Q_{l+1\|l+1})$	Uncertainty estimation
$F(\cdot)$	Expectation operator or covariance function.
P_{l+1}	Covariance matrix of process noise.
K	Identity or transition matrix (context: covariance update).
G	Measurement matrix mapping states to observation space.
L_{l+1}	Kalman gain at time step $l+1$.
W_{l+1}	Actual measurement value observed at time step $l+1$.
A	Raw data value before preprocessing (Z-score).
μ	Mean value of data.
σ	Standard deviation of data.
$E(a)$	ReLU activation function output.
a	Input to activation function or normalization.
Z_l	Convolution filter kernel weights.
y_{ji}	Bias term of the convolution filter.
$(g_l)_{ji}$	Output feature map value at pixel
j, i	Pixel indices in image/feature map.
l	Feature map index or layer index.
a^l	Normalized activation values in layer l.
$F(a^l)$	Mean of activations in layer lll.
$Var(a^l)$	Variance of activations in layer l.
ε	Small constant for numerical stability during normalization.
γ^l, β^l	Learnable scale and shift parameters in normalization.
$z1, g1, c1$	Input width, input height, and input depth respectively.
$z2, g2, c2$	Output width, output height, and output depth respectively.
e	Size of the pooling filter (e.g., 2×2).
X	Stride value for pooling or convolution.
TTT	Output size after pooling.
Z_i	Weights corresponding to class iii.
y	Weight vector associated with softmax.
$exp(\cdot)$	Exponential function used in softmax.
N	Number of output classes.
$Q'(a)$	Ground-truth distribution (one-hot encoded label).
$CrossEntropy$	Loss function for classification.

Dual-Channel Convolutional Neural Network with Mask Guidance for Digital Media Image Super-Resolution

Qian Song
School of Culture and Communication, Zibo 255300, China
E-mail: zibosongqian@163.com

Keywords: image design, reconstruction, digital media, image resolution, convolutional neural network, super resolution

Received: July 18, 2025

To address the issues of low resolution and detail loss in digital media visual image reconstruction, this paper proposes a super-resolution reconstruction algorithm based on a dual-channel mask-guided convolutional neural network (CNN). This algorithm innovatively designs a parallel guidance channel that utilizes edge masks generated from the input image as prior knowledge to enhance the main channel's ability to extract high-frequency structural features. Through multi-layer feature fusion and improved Adam optimizer, the model achieves a balance between reconstruction quality and efficiency. Tested on the CIFAR-10 and ILSVRC2020 datasets, and validated using image quality assessment metrics, it was found that under super-resolution tasks, the peak signal-to-noise ratio of the research method improved by an average of about 1.1 dB, with a maximum of 37.12 dB. The structural similarity index improved by an average of about 2.1%, with a maximum of 0.9682. At the same time, the algorithm demonstrated excellent efficiency advantages while maintaining high performance, with model parameters reduced to 1.2M, which was only 1/35 of advanced models in the same category. The average inference time for a single image has been reduced to about 20 ms, which was about 45% faster than the baseline method and demonstrated superior overall performance. The improved algorithm offers improved reconstruction quality and achieves a balance between quality and efficiency when the number of parameters is limited. Improving CNN through dual channel collaborative design and feature reuse mechanisms can provide new technological solutions for digital media image processing.

Povzetek: Predlagan je učinkovit dvo-kanalni CNN algoritem za super-ločljivost slik, ki z uporabo robnih mask izboljša kakovost rekonstrukcije ob hkratnem zmanjšanju števila parametrov in časa izvajanja.

1 Introduction

In recent years, digital media (DM) has emerged as a prominent focus of research due to its increasing applications in communication, visualization, and technology-driven solutions. The enhancement and reconstruction of visual images in the DM domain are key research areas, with a focus on improving image quality, resolution, and interpretability. The purpose is to develop methodologies for designing and reconstructing DM visuals to achieve higher realism and fidelity [1]. With advancements in relevant technology, image generation (IG) has found extensive applications in computer vision, including virtual reality (VR), augmented reality (AR), and automated image synthesis [2]. At present, there are many methods for designing and reconstructing DM images. However, high computational complexity may occur during the operation of these methods. This is mainly due to extensive parameter tuning and the need for large datasets. These issues lead to longer training times, and issues such as low image resolution often result in suboptimal reconstruction quality, limiting the practicality of these methods [3]. The development of current visual interactive images is hindered by factors such as low resolution, dynamic lighting changes, occlusion, and scale variations. Traditional super-resolution reconstruction algorithms primarily rely on either frequency domain or spatial domain features, with limited integration of both [4-5]. The current DM visual image reconstruction faces three major technological bottlenecks: 1. Insufficient feature utilization: Traditional super-resolution algorithms, such as interpolation and sparse coding, rely on artificially designed features and are difficult to adaptively extract multi-scale semantic information (such as edges and textures) from images. This results in problems such as edge blurring and loss of detail that often appear in the reconstruction results. 2. Conventional machine learning-based super-resolution reconstruction algorithms have disadvantages such as large data volume, complex computation, and inability to utilize some prior features of images, making it difficult to meet real-time requirements. 3. Lack of prior knowledge: The existing methods do not fully utilize the inherent features of the image, such as occluded edges, resulting in insufficient robustness to occlusion and dynamic lighting [6-7].

In response to the above challenges, various deep learning architectures have emerged in recent years. U-Net achieves multi-scale feature fusion through encoder-decoder structure and skip connections, and performs excellently in image reconstruction tasks [8]. Generative adversarial networks (GANs) enhance visual realism through adversarial training and are widely used in perception-oriented super-resolution reconstruction [9]. However, the complex structure of U-Net may result in high computational overhead, while the training of GAN is unstable and prone to introducing artifacts. In addition, existing methods rely heavily on network self-learning features and lack explicit guidance on image structure priors, which limits the accuracy in detail restoration and noise balance. This study aims to solve the above technical bottlenecks and meet the urgent demand for high-fidelity image reconstruction technology in digital imaging applications such as VR/AR and intelligent film and television production. It attempts to improve the quality of image feature fusion and reconstruction, and release the potential of hybrid modeling technology that combines data-driven and existing knowledge. Therefore, on the basis of existing research, this paper takes advantage of the super-resolution convolutional neural network (SRCNN) and proposes an improved CNN image super-resolution reconstruction algorithm under dual-channel-guided convolution. Moreover, guidance channels and mask images are introduced to enhance feature extraction capabilities. The innovation of this study is reflected in two aspects: First, the mask-guided dual-channel architecture. That is, based on the classic SRCNN, a second channel is introduced to specifically input the mask image generated by the residual transformation. It explicitly enhances high-frequency detail reconstruction through an adaptive feature fusion mechanism and prior edges to achieve a balance between noise reduction and detail preservation. The second is a lightweight design architecture. That is, the local structure is enhanced through the guidance channel, avoiding complex encoding and decoding or adversarial training, and improving the reconstruction efficiency and stability while maintaining a low number of parameters.

The research content is mainly divided into four sections. The subsequent section is organized as the literature section, which reviews the classic methods of image super-resolution reconstruction and related deep learning methods. The third section introduces the SRCNN reconstruction method and improvement ideas, including the addition of a guidance channel. The fourth section introduces the experimental design details and reconstruction performance of the improved CNN algorithm. The final section introduces the conclusions and future scope of the current research, as well as research prospects.

2 Related works

DM has gained traction in recent years due to its wide range of applications in communication, entertainment, and computational imaging. Wang et al. proposed a multi-scale extended CNN framework for image compression sensing and reconstruction by jointly training a fully convolutional structure measurement network and a reconstruction network. The results show that it outperforms existing methods in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [10]. Chen et al. proposed a two-stage image restoration (IR) network based on a parallel network and contextual attention to solve problems such as blurring and texture distortion in infrared. The results showed that this method could achieve a more realistic reduction effect [11]. Nakamura T utilized wave-optical deconvolution filters and color channel image synthesis to refine image reconstruction. The results showed that the research method overcame the traditional limitations and effectively improved the image resolution [12]. Pan X et al. proposed a transformer-based fully connected neural network for mask-based frameless image reconstruction, which improves the reconstruction ability through global feature inference. The results showed that it was superior to model-based and fully convolutional network methods [13]. Low-resolution images often resulted in loss of fine-grained details due to insufficient sampling rate and compression artifacts, and Chen et al. suggested that the low-frequency and high-frequency components of feature maps were treated equally in existing image super-resolution reconstruction methods [14]. Low-resolution images were difficult to accurately depict complex surface structures, and the complexity of spatial variations and structural features made it difficult to restore image structures, especially in some photometric stereoscopic images [15]. Yang et al. believed that sparse structural similarity could be used to evaluate the quality of distorted images, thereby improving the evaluation of image quality by ignoring the shortcomings of image structure. The results showed that the Pearson correlation coefficient of this method was 0.929, which could effectively improve the objectivity and accuracy of image quality assessment [16]. Low-resolution images often fail to provide sufficient data for analysis when containing scene-related information such as lighting, shadows, and reflections, thereby limiting the ability to extract effective information from the image. The challenges faced by low image resolution are particularly evident in image processing and reconstruction. Improving image quality through intelligent technology and next-generation super-resolution algorithms remains a key research focus, especially in addressing dynamic environments and hardware limitations.

At the same time, CNN image research is relatively mature. Zhang et al. proposed an infrared and visible image fusion algorithm based on ResNet-152 to extract multi-layer features by decomposing the low-frequency and high-frequency parts of the image. Experiments showed that the proposed method was superior to the comparison algorithm in retaining important features and obtaining more details [17]. Shao G et al. proposed a subpixel CNN for image super-resolution reconstruction, converting RGB to YCbCr mode, and introducing residual networks and upsampled subpixel convolutional layers (CLs) for improvement. Experimental results showed that it outperformed many traditional methods in terms of

accuracy and time consumption [18]. Jin Z proposed a flexible deep CNN framework for processing in computer vision, which utilized the frequency characteristics of different types of workpieces. By adjusting the architecture, the same method could be utilized for various IR tasks. The results illustrated that the framework was more excellent than other methods [19]. Zhang Y designed a universal image fusion model for CNN. This study first utilized two CLs to extract image features from the II. Then, the features of the II were fused by selecting appropriate fusion rules. The results showed that compared with other fusion models, the research model had better pan-Chinese ability and achieved better fusion results [20]. Zhu F et al. proposed a CNN-based denoising method for better obtaining clean images from noise. Combining some different rate expansion convolutions with common convolutions enriched the features extracted from multi-layer convolutions. The results showed that the denoising effect of the research method was excellent [21]. To improve the edge detection performance of noisy images, Yuan S et al. proposed an edge detection method based on nonlinear structural tensors, which determined image edges by calculating the tensor product, eigenvalues, and eigenvectors of noisy images. The results indicated that this method could detect more monitoring points compared to other methods, with a shorter average detection time and overall better detection performance [22].

In summary, DM has brought convenience to people's lives. There is relatively abundant research on visual Image Design (ID) and image reconstruction, but there is relatively little research on CNNs. The development of CNN is relatively mature and has achieved good results in image denoising, feature extraction, and detection. In view of this, this study proposes a new CNN-based DM visual ID and image reconstruction method.

3 ID and image reconstruction method for CNN algorithm

To improve the reconstruction quality of DM visual images, a SRCNN dual-channel reconstruction algorithm is proposed, which adds guided channels and uses masked images for improvement. The improvement and innovation of this study is the utilization of explicit edge mask-guided parallel channels to provide powerful structural priors for super-resolution reconstruction. Its mechanism is essentially different from traditional implicit feature learning or simple multi-scale fusion, and can greatly improve the quality of image reconstruction.

3.1 SRCNN reconstruction algorithm

There are various methods for image super-resolution, including instance-based, interpolation-based, and reconstruction-based methods and deep learning algorithms [23]. In CNN algorithms, SRCNN is a representative algorithm applied to image super-resolution, which can learn the unique connections between high-resolution and low-resolution images and use these connections as constraints to generate high-resolution images. Research has shown that using multi-layer network structures can enable models to learn more complex features and patterns, which is particularly important for generating high-quality images and reconstruction tasks [24]. Figure 1 shows the relevant framework.

As shown in Figure 1, the three CLs are utilized for feature extraction, nonlinear mapping, and high-resolution image reconstruction. The entire process is to first extract small blocks from the interpolated low-resolution image and represent the extracted small blocks in the form of high-dimensional vectors. Then, it transforms all high-dimensional vectors nonlinearly and maps them onto other high-dimensional vectors [25]. Finally, it aggregates all high-resolution small blocks to form the final high-resolution image, which should be similar to the ideal high-resolution image corresponding to the low-resolution image. The SRCNN algorithm mainly uses convolution operations to train and process input images during forward and backward propagation. The CL can convert digital signals into low-dimensional vectors and store the extracted features in the feature vectors. The convergence rate of the algorithm is restricted by the selection of the Loss function. This study uses the fast convergence cross-entropy as the loss function and adds a minimum value to it, as shown in equation (1).

$$\begin{cases} \hat{y}' = \hat{y} + \sigma \\ C = -\left[y \ln \hat{y}' + (1-y)\ln(1-\hat{y}') \right] \end{cases} \tag{1}$$

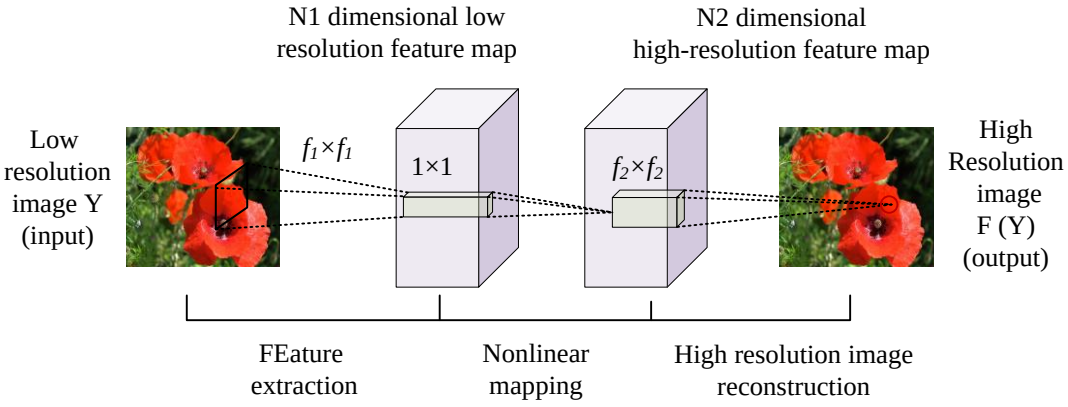


Figure 1: The network framework of SRCNN.

As shown in equation (1), the predicted output is represented by $\hat{y}$. The actual output is represented by y. σ represents the minimum value. The commonly used CNN algorithm is aimed at the gradient descent method. Although it has good performance, it also has disadvantages, such as difficulty in choosing an appropriate learning rate, and easy occurrence of saddle points and local extremum problems [26]. Therefore, an adaptive learning rate is selected to overcome the shortcomings. The Adam algorithm is selected in this study. The Adam algorithm can dynamically adjust the learning rate of each parameter, and it can retain historical gradient information, accelerate the convergence process, and reduce oscillation phenomena. The random gradient descent algorithm for the Adam algorithm is shown in equation (2).

$$\begin{cases} m_t = \rho_1 * m_{t-1} + (1-\rho_1) * g_t \\ n_t = \rho_2 * n_{t-1} + (1-\rho_2) * g_t^{\,2} \\ \hat{m}_t = \dfrac{m_t}{1-\rho_1^t} \\ \hat{n}_t = \dfrac{n_t}{1-\rho_1^t} \\ \Delta\theta = -\dfrac{\hat{m}_t}{\sqrt{\hat{n}_t} + \delta} * \eta \\ \theta_{t+1} = \theta_t + \Delta\theta \end{cases} \tag{2}$$

In equation (2), the gradient is represented by g_t. ρ_1 and ρ_2 are the decay rates. m is the first-order moment of the gradient. $\hat{m}_t$ is the moment estimate after deviation correction of the first-order moment of the gradient. n_t is the second-order moment of the gradient. $\hat{n}_t$ is the moment estimate after deviation correction of the second-order moment of the gradient. Learning rate is expressed in η. δ represents the minimum constant. A large number of pooling operations may result in blurry feature information. Therefore, a *soft* max function classifier is added after the fully connected layer to convert complex result values into relative probabilities that are easy to compare and understand. There are a large number of shared layers between different neurons, which can effectively reduce the amount of data. Research has been conducted to improve traditional neurons by using random deactivation, as shown in equation (3).

$$\begin{cases} r^{(l)} \sim Bernoulli(p) \\ \tilde{a}^{(l)} = r^{(l)} * a^{(l)} \\ c^{(l+1)} = w^{(l+1)} \tilde{a}^{l} + b^{(l+1)} \\ a^{(l+1)} = f(c^{l+1}) \end{cases} \tag{3}$$

As shown in equation (3), the selected neuron is represented by $\tilde{a}^{(l)}$; The sample subset of Bernoulli distribution satisfying probability p is represented by $r^{(l)}$. $a^{(l+1)}$ represents $l+1$ layer output. $l+1$ input weighting and vector are represented by $c^{(l+1)}$. The weight of $l+1$-layer is represented by $w^{(l+1)}$. $b^{(l+1)}$ is the $l+1$-layer bias. $f()$ is the excitation function.

3.2 Design of Improved SRCNN Dual-Channel Mask-Guided Network Architecture Based on Image Reconstruction

3.2.1 Overall Network Architecture

Although SRCNN has good recognition and image reconstruction performance, its network structure relies heavily on its own neural network structure, resulting in limited learning ability. Therefore, to improve SRCNN, this study introduces a parallel guidance channel and utilizes edge masks explicitly extracted from the input image as structural priors to guide the feature learning process of the main channel. The purpose is to enhance the network's ability to recover high-frequency details and further improve the reconstruction performance of the neural network. The improved algorithm structure is shown in Figure 2.

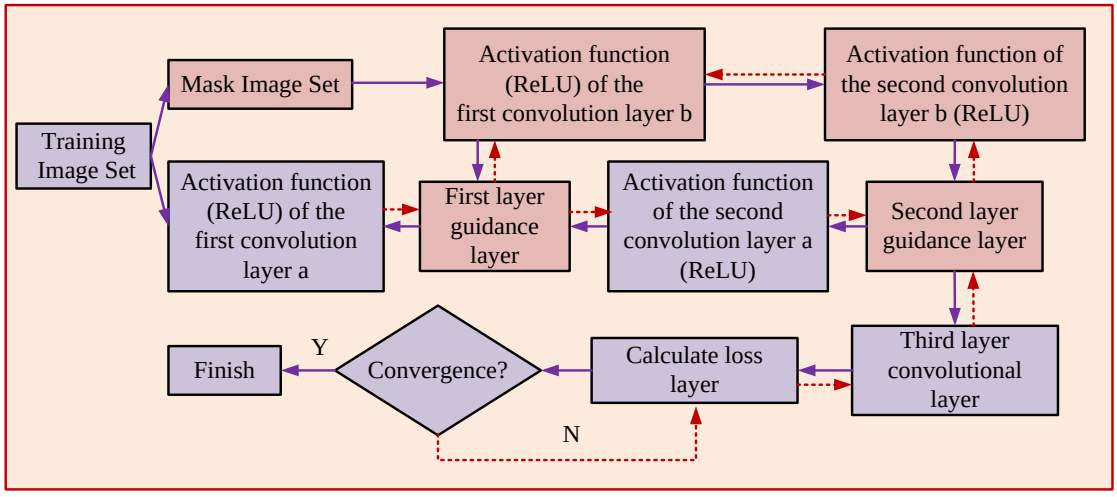


Figure 2: Improved SRCNN algorithm structure.

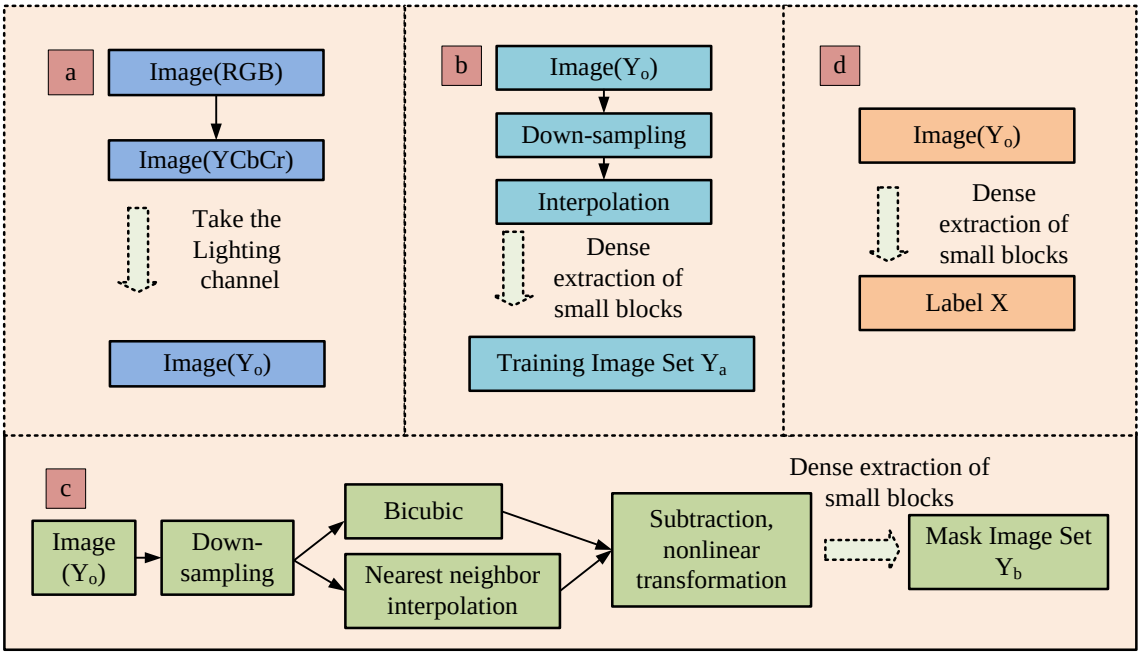


Figure 3: The process of obtaining a training image set.

As shown in Figure 2, the improved network has three CLs and two channels. The solid line represents forward conduction, while the red dashed line represents reverse conduction. The red part represents the improved part, which is the newly added second channel. After adding channels, five modules are added, namely the first guiding layer, the second guiding layer, the first CL, the second CL, and the mask image dataset. By introducing additional guidance channels and using masked images of input images, new dimensions can be added to feature learning, thereby improving network efficiency. Specifically, the guidance channel can provide additional information at different scales related to the main input image, thereby enhancing the network's perception ability of different features. Mask images are used to emphasize important features such as edges, textures, and structures in the image, and utilize these features to guide the original convolutional channels.

3.2.2 Mask generation module

To provide clear edge guidance information to the network, a corresponding mask image is first generated for each input image. The specific method is to use different interpolation algorithms to address the differences in edge processing characteristics. Firstly, the original low-resolution image is interpolated using bicubic interpolation and nearest neighbor interpolation to obtain two images, which are then subtracted and subjected to a nonlinear transformation to enhance the edges. The image training method can be specifically shown in Figure 3.

In Figure 3, firstly, the images obtained from the training set are transferred from the RGB color space to the YCbCr color space, and the image Y0 is obtained through the illumination channel. During image downsampling, after double triple interpolation, overlapping small blocks are extracted from the interpolated image, and then the extracted small blocks are aggregated to form a training image set Y_a. Bicubic interpolation and nearest neighbor interpolation are used to upsample the original low-resolution image, and then the two resulting images are subtracted and the edges of the image are enhanced through nonlinear transformation. This process can be represented by equation (4).

$$Y_{mask} = \tanh(\beta * | f_b(Y_{ds}) - f_n(Y_{ds}) |) \quad (4)$$

As shown in equation (4), the mask image is represented by Y_{mask}. The sampled image is represented by Y_{ds}. Bicubic interpolation is represented by f_b. The nearest neighbor interpolation is represented by f_n. The transformation size coefficient of nonlinear transformation is represented by β, which is generally negative. tanh() is the hyperbolic tangent activation function, which normalizes the mask values between (-1,1) to make the edge features more prominent. By performing differential operations on bicubic interpolation (preserving smooth regions) and nearest neighbor interpolation (highlighting edge step effects), high-frequency edge information in the image can be enhanced, mask image M can be generated, and the network can be guided to focus on structural features. Traditional SRCNN relies solely on CL learning for low-resolution to high-resolution mapping, lacking explicit utilization of inherent image features such as edges and textures. The addition of a guiding channel can optimize the input mask image through dual-channel collaboration, preserving its details and improving its image reconstruction effect. The acquisition of labels is shown in Figure 4.

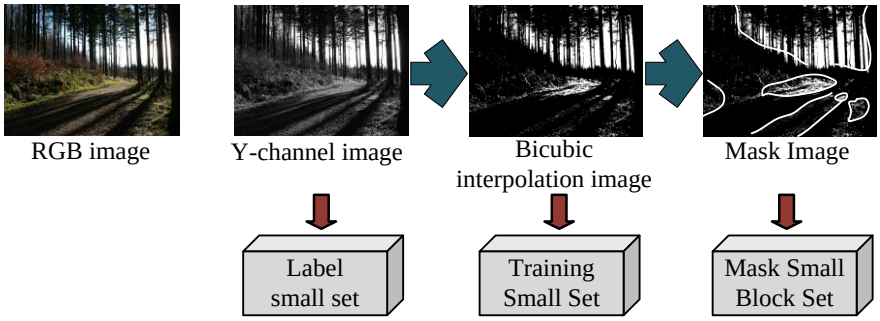


Figure 4: Acquisition of labels.

Figure 4 shows the specific extraction process of label X. This is a randomly selected image from the training set, displaying mask image blocks, training image blocks, and label blocks, respectively. The purpose of the first CL is feature extraction, extracting small blocks from images Y_a and Y_b to obtain II feature $F_1(Y)$, which is similar to the case through a set of filters, and its expression is shown in equation (5).

$$\begin{cases} F_{1a}(Y) = \max(0, W_{1a} * Y_a + B_{1a}) \\ F_{1b}(Y) = \max(0, W_{1b} * Y_b + B_{1b}) \end{cases} \quad (5)$$

As shown in equation (5), the two channels are denoted by a and b. $F_{1a}(Y)$ is the content feature map output by the main channel in the first layer, and $F_{1b}(Y)$ is the structural feature map output by the guiding channel in the first layer. The convolution kernels (CKs) of the two channels are represented by W_{1a} and W_{1b}, respectively. Y_a and Y_b represent training images and mask images, respectively. $*$ represents convolution operation. The bias vectors of the two channels are represented by B_{1a} and B_{1b}

3.2.3 Dual-channel feature extraction and fusion

After the mask is generated, the upsampled image and the mask image are fed into the main and guide channels, respectively. The main channel is responsible for learning the overall content and texture information of the image, and the guide channel is responsible for learning the structure and edge information of the image. The dual-channel separation design can avoid feature confusion and provide prior structural constraints. The purpose of the second layer of CL is nonlinear mapping, that is, to transform the feature vector of low-resolution space into high-resolution space. At this time, the output feature is $F_2(Y)$, and its expression is shown in equation (6).

$$\begin{cases} F_{2a}(Y) = \max(0, W_{2a} * G_1(Y) + B_{2a}) \\ F_{2b}(Y) = \max(0, W_{2b} * F_{1b}(Y) + B_{2b}) \end{cases} \quad (6)$$

As shown in equation (6), the CK of two channels are represented by W_{2a} and W_{2b}, respectively. The output of the first guidance layer is represented by $G_1(Y)$. The output of the first CL is represented by $F_{1b}(Y)$. The bias vectors of the two channels are represented by B_{2a} and B_{2b}. The nonlinear mapping of the second layer convolution can map low-dimensional features to high-dimensional space and fuse dual-channel information. Guide channel features can enhance edge consistency and prevent reconstruction blurring. The purpose of the third layer convolution is to reconstruct high-resolution image. The final generated high-resolution image is $F(Y)$, and its expression is shown in equation (7).

$$F(Y) = W_3 * G_2(Y) + B_3 \quad (7)$$

As shown in equation (7), the CK is represented by W_3. The output of the second guidance layer is represented by $G_2(Y)$. The bias vector is represented by B_3. The third layer does not need to undergo nonlinear transformation, as the W_3 of the third layer can be regarded as a filter, and the reconstruction process can be considered as a linear transformation process. The mask image in Figure 4 shows that it can represent edge features, and these features are used to guide the convolutional channel. The formula for the guidance layer is shown in equation (8).

$$G_i(Y) = F_{ia}(Y) \times F_{ib}(Y) \quad i = 1, 2 \quad (8)$$

As shown in equation (8), $F_i(Y)$ represents the output image of the i-th CL. The research method requires training parameters $\{W_{1a}, W_{1b}, W_{2a}, W_{2b}, W_3, B_{1b}, B_{2b}, B_3\}$, and the training process is actually a part of parameter optimization and estimation. By reducing the loss between high-resolution image and reconstructed images, the optimal solution of the parameters is found. The visualization of guide layer features validates the effective utilization of edge information. To train the network parameters, the study aims to optimize the model by minimizing the mean square error (MSE) between the reconstructed image and the real high-resolution image, and its expression is shown in equation (9).

$$L(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \left\| F(Y_i; \Theta X_i) \right\|^2 \quad (9)$$

As shown in equation (10), the reconstructed image is represented by $F(Y_i; \Theta X_i)$. The number of training samples is represented by n. The high-resolution image set is represented by $\{X_i\}$. The set of processed images with the same size as the original image is represented by $\{Y_i\}$.

4 Result analysis of ID and image reconstruction methods for CNN algorithm

4.1 Experimental design platform and related parameter settings

To evaluate the proposed CNN-based DM visual ID and image reconstruction method, multiple IG tasks are selected in this experiment. It includes facial recognition, facial IR, IG, and IR, with multiple publicly available datasets. This study first selects the parameters of the CNN and then tests and verifies the performance of the algorithm. To avoid experimental errors caused by different equipment, the study uses the same computer for the experiment, using Intel Core i7-9700 3.00GHz CPU and 20GB memory. Operating System: Ubuntu 20.04 LTS, with the deep learning framework of PyTorch 1.10. The core library is CUDA 11.3, Python 3.8, and OpenCV 4.5. The experiment uniformly adjusts the input image to 256 × 256 pixels (based on common preprocessing standards such as CIFAR-10 and ILSVRC2020) to meet the requirements of multi-scale feature extraction. The size of the first CK is 9×9 (large kernel captures global structure), the size of the second CK is 6×6 (moderate kernel balances details and computational complexity), and the size of the third CK is 5×5 (small kernel refines local features). Stride defaults to 1 to ensure high-resolution feature preservation (applicable to all CLs). In addition, in the second layer of the guidance channel, Stride=2 is used to downsample the mask image to reduce computational complexity. The AF of the first- and second-layer channels is LeakyReLU (α=0.2), which alleviates the problem of gradient vanishing. The third layer channel is linearly activated (without AF), as the reconstruction task requires retaining all intensity values. ReLU is used throughout the guidance channel to enhance the sparsity of edge features. The initial learning rate is 0.0001, and the optimizer is Adam (β_1=0.9, β_2=0.999, and ε=1e-8). Image blocks of 96*96 pixels are randomly cropped from the high-resolution image. The corresponding low-resolution image block is obtained by performing bicubic downsampling (magnification factor x3) on the high-resolution image block. To increase data diversity and improve model generalization ability, online data augmentation is applied to the training data, including random horizontal flipping and 90-degree rotation. The optimizer uses improved Adam, and the initial learning rate is set to 1e-4. After every 20 rounds, the learning rate decays to 0.5 times the original value. Table 1 shows the experimental parameter settings.

4.2 Dataset source and related information

The CIFAR-10 dataset and ILSVRC2020 dataset have a wide range of applications and can be used to evaluate image processing and reconstruction algorithms. The CIFAR-10 dataset contains a variety of color image categories and has rich visual effects, which can effectively test the algorithm's processing effect on visual images. The ILSVRC2020 dataset has a large amount of data, rich sample size, and high image resolution, making it more suitable for complex scenes and high detail testing content. The use of CIFAR-10 and ILSVRC2020 datasets to evaluate the performance of ID and image reconstruction algorithms is due to their widespread application and recognized benchmark position in the field of computer vision. Moreover, their diversity, rich content, differences in data volume and structure, as well as balanced data distribution, facilitate the comprehensive evaluation of algorithm performance and robustness from multiple aspects. This study selects 100 images from the CIFAR-10 and ILSVRC2020 datasets for testing experiments. 100 images are randomly selected from CIFAR-10 and ILSVRC2020 as the test set, and the remaining data are used for training. The extraction of the test set adopts stratified sampling, where CIFAR-10 is uniformly selected into 10 categories, with 10 images per category. ILSVRC2020 is grouped according to the semantic hierarchy of ImageNet (such as animals and artificial objects), with 20 images extracted from each group. All experiments are repeated 5 times, with random seeds fixed at {42, 123, 2023, 55, 17}, and the results are taken as mean ± standard deviation.

4.3 Parameter experiment of CNN network structure

For selecting an appropriate size of CK, the interpolation ratio is 3, the learning rate is 0.0001, and the quantity of CK in the first layer is 64. The quantity of CK in the second layer is 32, and an unimproved CNN algorithm is introduced as a reference to compare the loss values (LVs) of reconstructed images with CK sizes of 9-2-5, 9-4-5, and 9-6-5, respectively. The results are shown in Figure 5.

Table 1: Experimental parameter settings.

Parameter	Numerical value
Batch Size	16
Epochs	100
Loss function	MSE + 0.5×MSSIM
Weight initialization	He follows a normal distribution
Decline in learning rate	2
CK size	9-6-5
Guide channel downsampling Stride	2 (second layer only)

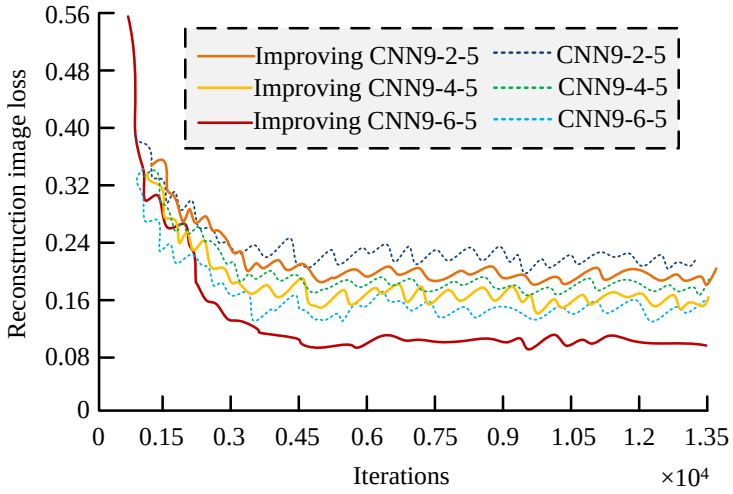


Figure 5: Reconstruction image LV for different CK sizes.

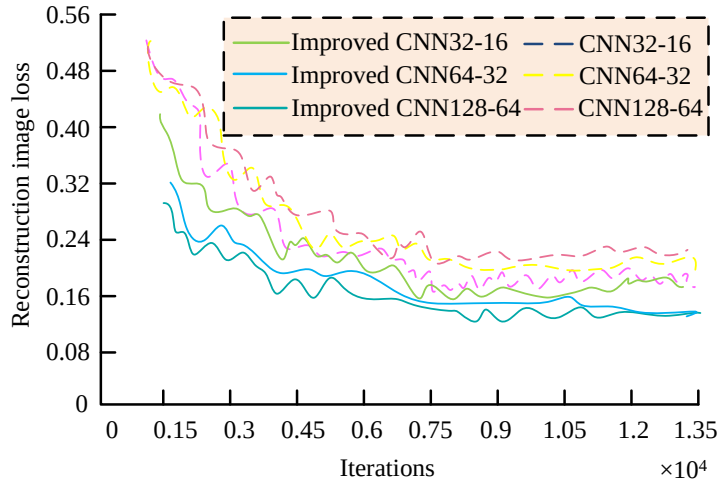


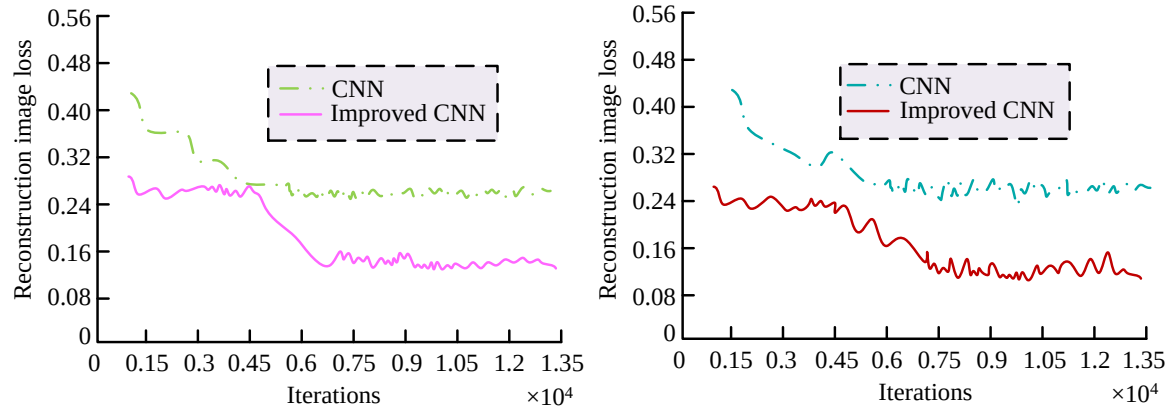
Figure 6: Loss of reconstructed images with different CK numbers.

Figure 5 shows the loss result graph plotted before 13,500 iterations. Compared with the unmodified CNN, the improved CNN has a significantly faster rate of convergence. At the iteration number of 4,000, the CNN of each size has a significant convergence, and the LVs are 0.10, 0.16, and 0.20 respectively, while the unmodified CNN of each size has an LV of 0.15, 0.18, and 0.24, and does not converge to the minimum value. Figure 5 also indicates that as the size of the second CK increases, the LV of the reconstructed image gradually decreases. This indicates that an increase in the size of the second CK will enable the neural network structure to learn more information, reduce LV, and achieve better ID and image reconstruction results. The number of CK also affects the experimental results. Choosing the same parameters as above, the fixed CK size is 9-6-5, the quantity of CK is 128 in the first layer, 64 in the second layer, 64 in the first layer, 32 in the second layer, 32 in the first layer, and 16 in the second layer. The LV of reconstructed images with different CK numbers is shown in Figure 6.

As shown in Figure 6, as the number of CK increases, the neural network structure learns more information, the LV of the reconstructed image is lower, and the image effect is better. Compared to the unmodified CNN, the improved CNN converges earlier. When the quantity of convolutions in the first layer is 128, the quantity of convolutions in the second layer is 64, and the quantity of iterations is 7,000, the improved CNN convergence LV is 0.15. At this point, the LV of the unimproved CNN is 0.23 and has not yet converged to the minimum value.

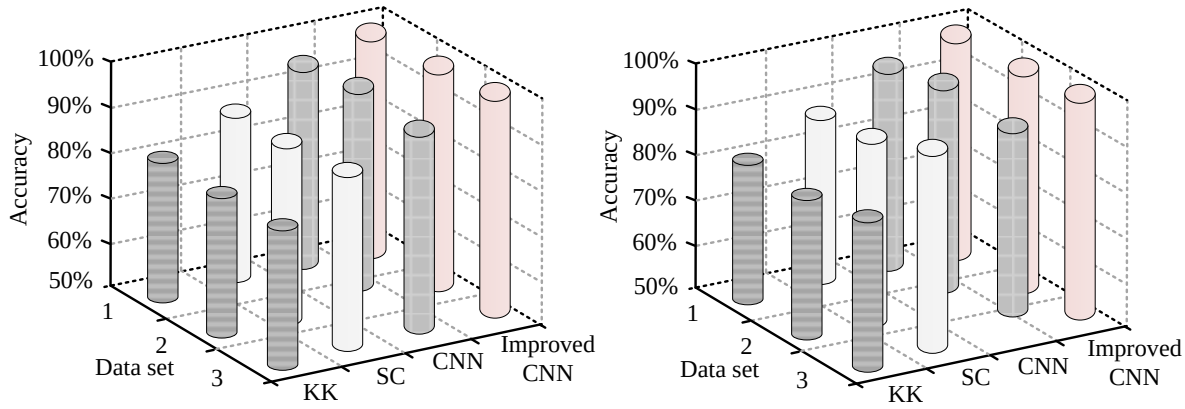
4.4 Comparative analysis experiment of ID and image reconstruction methods for CNN

To verify the superiority of DM visual ID and image reconstruction methods over CNN, an unimproved CNN algorithm is further introduced for comparison. At this point, the CK size is set to 9-6-5 and the quantity of CK in the first layer is 128, while the quantity of CK in the second layer is 64. The comparison results are shown in Figure 7.



(a) CIFAR-10 (b) ILSVRC2020

Figure 7: Comparative experiments of two datasets.



(a) CIFAR-10 (b) ILSVRC2020

Figure 8: Comparison of accuracy of four methods.

Figure 7 (a) shows that in the dataset CIFAR-10, CNN converges at 4,500, leading to an LV of 0.25. The improved CNN converges at 7,500, leading to an LV of 0.16. Figure 7 (b) shows that in the dataset ILSVRC2020, CNN converges at 6,000, leading to an LV of 0.29. The improved CNN converges at 7,000, leading to an LV of 0.10. This indicates that the improved research method results in lower LV and better results in obtaining the final design and reconstruction images. To further validate the superiority of the research algorithm, image reconstruction methods are introduced: sparse regression and natural image prior (KK) and sparse coding-based method (SC) are used to compare the accuracy, error, response time, and image evaluation results of the four methods. In the CIFAR-10 dataset, images are divided into high, medium, and low layers based on their edge complexity (gradient amplitude calculated using the Sobel operator). 30 images are randomly selected from each layer to a ensure balanced distribution of structural complexity among the three groups. The accuracy of the four methods is shown in Figure 8.

Figure 8 shows that the improved CNN is superior to the other three algorithms in accuracy on both datasets. The accuracy of the algorithm is ranked from high to low as Improved CNN, CNN, SC, and KK. In Figure 8 (a), the improved CNN method achieves the highest accuracy of 97.23% and an average of 95.23% for the four sets of data. In Figure 8 (b), the improved CNN method achieves the highest accuracy of 98.23% and an average of 96.25% for the four datasets. Overall, the accuracy of the improved CNN performs well on both datasets. The mean absolute percentage error (MAPE) of the four methods is shown in Figure 9.

In Figure 9(a), the MAPE of the 20 images of the SC and KK algorithms is 3.92% and 4.12%. The improved CNN curve fluctuated the most, with a MAPE of 1.02% for 20 images. In Figures 9(a) and (b), the improved CNN has minimal MAPE. The proposed method includes guide channel and mask images, which can effectively extract image edge features and enhance feature extraction capabilities. However, traditional sparse coding or regression methods are often based on fixed prior information and lack flexibility. Sparse regression and natural image priors often rely on linear combinations, which makes it difficult to capture complex features. The improved CNN has good accuracy, error rate and response time. Ten photos are selected from the two datasets for testing. The response times of the four methods are shown in Figure 10.

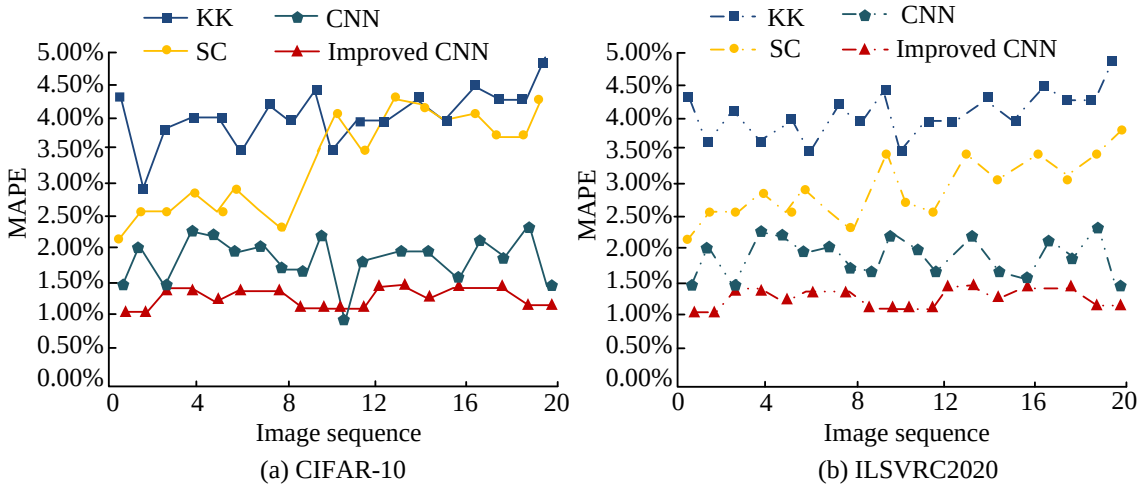


Figure 9: Comparison of MAPE of four models.

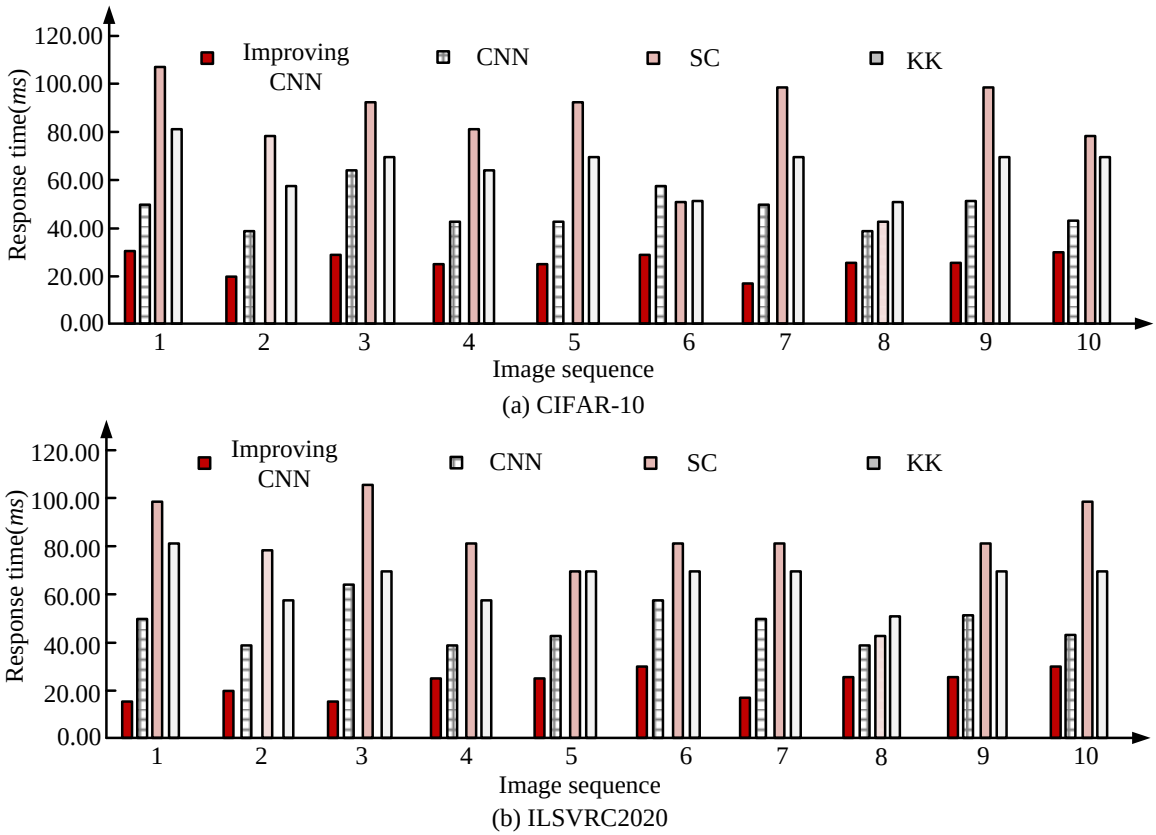


Figure 10: Comparison of response times of four methods.

Figure 10 shows that the response time of the improved CNN is lower than that of the other three algorithms. The response times of these algorithms from low to high are improved CNN, CNN, KK, and SC. Figure 10 (a) shows that the minimum response time of the improved CNN method is 18 *ms* and the average value is 21 *ms*. Figure 10 (b) shows that the minimum response time of the improved CNN method is 16 *ms*, and the average value is 20 *ms*. The improvement research scheme can coordinate the learning process of the guide channel and the main input channel more effectively by jointly optimizing the loss function of the guide, thereby significantly improving the reconstruction effect. The simple integration of the boot channel does not significantly increase the computational complexity, and its precise grasp of key information significantly enhances performance while maintaining responsiveness. To further analyze the computational efficiency of the research method, the performance differences between the research method and CNN, KK, and SC methods are compared in terms of parameter size, inference time, FLOPs, and other indicators. The results are shown in Table 2.

Table 2: Comparison of the computational efficiency of different algorithms.

Index	Parameter (M)	FLOPs (G)	GPU Memory (GB)	Inference time (ms)	PSNR (dB)
KK	0.27	0.8	0.5	35	28.50
SC	0.12	2.3	1.2	45	28.21
CNN	0.24	4.7	1.8	25	30.05
Improved CNN	0.38	3.5	1.6	20	32.52

Table 3: Comparison of image quality evaluation results of four methods.

Evaluation criterion	Interpolation Scale	Double triple interpolation	KK	SC	CNN	Improved CNN
MSSIM	2	0.9976	0.9932	-	0.9975	0.9981
	3	0.9738	0.9769	0.9801	0.9814	0.9875
	4	0.9517	0.9687	-	0.9712	0.9789
WPSNR	2	50.12	57.02	-	58.97	60.64
	3	41.19	46.87	47.12	47.19	47.99
	4	37.33	40.48	-	47.95	41.98
NQM	2	36.97	38.9	-	41.23	43.98
	3	27.75	32.87	33.04	33.17	34.54
	4	21.86	24.34	-	25.65	26.87
IFC	2	6.19	6.46	-	8.21	8.47
	3	3.53	3.89	4.57	4.79	5.01
	4	2.23	2.56	-	3.74	3.88
SSIM	2	0.9268	0.9521	-	0.9578	0.9682
	3	0.8648	0.8711	0.8864	0.9068	0.9369
	4	0.8148	0.8518	-	0.8647	0.8738
PSNR	2	33.21	36.89	-	36.98	37.12
	3	30.23	31.51	32.87	33.86	34.97
	4	28.75	30.48	-	30.89	31.97

In Table 2, the number of parameters (0.38M) of the proposed improved CNN is increased by 58% compared with the original CNN, which is mainly due to the CK of the new guide channel. The improved CNN can control the growth through the parameter sharing strategy (such as the mask channel shares some weights with the main channel). The FLOPs of the improved CNN are reduced by 25% (3.5G vs 4.7G), which benefits from the sparsity of mask guidance and skips the low-contribution region in the second layer convolution. The inference speed (20 ms) of the improved CNN is better than that of SC (45 ms) and CNN (25 ms), and the memory occupation (1.6GB) is between SRCNN (1.8GB) and SC (1.2GB). The improved CNN has better reconstruction quality (PSNR: 32.52 dB), which is 2.47 dB higher than that of the original SRCNN. KK is parameter-dependent, but has a low PSNR (28.50 dB) because it cannot learn data-driven features. SC needs iterative optimization, the inference speed is slow (45ms), and it is sensitive to occlusion. The improved CNN can achieve quality-efficiency balance under the condition that the number of parameters is limited. To further validate the applicability of the research method, image quality evaluation indicators such as PSNR, SSIM, information fidelity criterion (IFC), noise quality measure (NQM), weighted PSNR (WPSNR) are introduced. The multi-scale SSIM (MSSIM) evaluates image quality. The IFC considers the original image as the source of information and the distortion process as an information channel. It evaluates the image quality and the fidelity of information transmission by calculating how much information can be extracted from the reconstructed image. IFC has a strong correlation with human perception of the naturalness and clarity of images. NQM works by simulating the sensitivity of the human visual system to different spatial frequencies, effectively quantifying annoying distortions such as block effects and ringing effects introduced by compression or reconstruction algorithms. The lower the value of NQM, the less perceived noise and higher quality of the image. WPSNR assigns different weights to different regions when calculating errors, and better reflects the reconstruction quality of visually important regions than standard PSNR. Under the optimal parameter settings, 20 photos are randomly selected from the CIFAR-10 data for application, and the outcomes are indicated in Table 3. The 20 samples for image quality evaluation include 5 high texture, 5 low texture, 5 high dynamic range, and 5 occluded scenes.

In Table 3, the MSSIM value of the improved CNN is greater than 0.97 under different interpolation ratios, while the MSSIM value of the KK and CNN is slightly worse. The improved CNN achieved WPSNR values of 60.64, 60.64, and 41.98 at interpolation ratios of 2, 3, and 4, respectively, which was still higher than other comparison algorithms under the same conditions. The NQM values (43.98, 34.54, 26.87) of the improved CNN outperforms the other methods in all cases, and the IFC value is stable. The above results indicate that the proposed algorithm achieves higher similarity in image content reconstruction results at different scales, and performs better than other algorithms in feature extraction and image reconstruction. Among them, the SC performs the worst, as it is difficult to grasp image details and structural content solely through sparse encoding methods. The KK performs better than the SC, but its prior thinking based on natural images is difficult to ensure structural similarity and reconstruction results. To further evaluate the performance of the improved CNN based on dual-channel guidance, it is compared with the residual channel attention network (RCAN), enhanced deep super resolution (EDSR), super-resolution generative adversarial network (SRGAN), dense residual network (DRN), and Swin Transformer for Image restoration (SwinIR) on the CIFAR-10 dataset. The interpolation ratio

is 3, and 50 test images are randomly selected for quantitative evaluation. The comparison results are shown in Table 4.

In Table 4, SwinIR performs the best on most objective metrics (PSNR, SSIM, IFC, NQM, and WPSNR), thanks to the powerful global modeling capability of the Transformer architecture. The PSNR and SSIM of the improved CNN are only 0.15 dB and 0.0016 lower than SwinIR, proving its good reconstruction quality. Improved CNN outperforms all comparison methods in MSSIM metrics (0.9875), indicating its advantage in maintaining multi-scale structural similarity. Its parameter size is only 1.2M, far lower than complex networks such as EDSR (43.1M) and DRN (22.3M), demonstrating good parameter efficiency. Improved CNN's 21.3 ms inference time is about 45% faster than EDSR and about 66% faster than SwinIR, resulting in a significant reduction in computational cost. RCAN improves performance by channel attention, but has a large number of parameters (156 m). To further validate the improved CNN, it is compared with the Channel Spatial Hybrid Attention Network for Image Super Resolution (CSHA), Multi-level Information Compensated U-Net for Image Super Resolution (MICHNet), SRGAN, and Soft Edge-guided Progressive Super Resolution Network (SEPNet) in terms of computational efficiency, reconstruction quality, resource consumption, and reconstruction accuracy. The comparison results are shown in Table 5.

In the results of Table 5, the research method maximizes the preservation of the structural information of the original image due to the dual-channel design (main channel+mask guidance), resulting in the best information fidelity (8.47) compared to CSHA (7.82), MICHNet (7.35), SRGAN (6.21), and SEPNet (8.12). Although SEPNet adopts edge compensation strategy (IFC=8.12), it does not explicitly model the inter channel dependencies, which is slightly inferior to the research method. The NQM value of the research method (43.98) performs well, increasing by 9.6% compared to CSHA (40.12), while SRGAN has the lowest NQM value (35.44) due to the introduction of high-frequency artifacts in adversarial training. The WPSNR value of the research method (60.64 dB) verifies the effectiveness of mask guidance, while MICHNet (56.91 dB) performs poorly in areas with complex textures due to feature loss issues. The MAPE of the research method is the lowest (1.00%), a decrease of 10.7% compared to SEPNet (1.12%), while the MAPE of SRGAN reaches 2.32%, mainly due to systematic biases in the generated images (such as color shift). The reconstruction quality of the research method performs well, with an SSIM value (32.52 dB), 0.72 dB higher than CSHA and significantly better than SRGAN (+2.32 dB). The reason for this is that the improved CNN explicitly enhances edge features through mask-guided channels, while CSHA only indirectly allocates resources through attention weights. SRGAN is prone to artifacts due to GAN. The above structure indicates that the improved CNN is comprehensively leading in perception indicators such as IFC, NQM, and WPSNR. This indicates that the improved CNN can improve pixel accuracy while also better maintaining visual naturalness and balancing computational efficiency to a certain extent. Afterwards, the image reconstruction effects of the above comparison methods are presented, and the results are shown in Figure 11.

The results in Figure 11 indicate that the image clarity trained by the improved CNN is significantly higher than other comparison models, with higher image clarity and better restoration of image details. Secondly, the SEPNet and CSHA perform well. Afterwards, the comparison results of different algorithms are analyzed using the statistical tests in Table 6.

Table 4: Image quality evaluation results of different comparison methods.

Evaluation metric	RCAN (ECCV'18)	EDSR (CVPR'17)	SRGAN (CVPR'17)	DRN (CVPR'18)	SwinIR (ICCV'21)	Improved CNN
PSNR (dB)	34.21	33.98	32.45	34.35	35.12	34.97
SSIM	0.9321	0.9287	0.9123	0.9338	0.9385	0.9369
IFC	4.89	4.76	4.21	4.95	5.12	5.01
NQM	33.87	33.45	31.89	34.02	34.78	34.54
WPSNR (dB)	47.45	47.12	45.89	47.56	48.21	47.99
MSSSIM	0.9845	0.9832	0.9789	0.9851	0.9867	0.9875
Parameter quantity (M)	15.6	43.1	1.5	22.3	11.9	1.2
Inference time (ms)	45.2	38.7	52.3	41.5	62.8	21.3
Training time (m)	48	36	72	54	96	24

Table 5: Comparison of reconstruction performance of different algorithms.

Model	PSNR (dB)	SSIM	IFC	NQM	WPSNR (dB)	MAPE (%)	FLOPs (G)	Inference time (ms)	GPU Memory (GB)
Improved CNN	32.52	0.9682	8.47	43.98	60.64	1.00	3.5	20	1.6
CSHA	31.80	0.9621	7.82	40.12	58.23	1.45	8.7	35	2.8
MICUNet	31.45	0.9583	7.35	38.76	56.91	1.78	12.4	50	3.5
SRGAN	30.20	0.9456	6.21	35.44	53.89	2.32	22.9	80	5.2
SEPNet	32.10	0.9650	8.12	42.05	59.87	1.12	9.2	45	2.5



Figure 11: Image reconstruction effects of different algorithms.

Table 6: Statistical test results.

Index	Improved CNN vs CSHA	Improved CNN vs MICUNet	Improved CNN vs SRGAN	Improved CNN vs SEPNet
PSNR (dB)	t=4.32** d=1.25	t=5.67*** d=1.87	t=8.91*** d=3.02	t=2.15* d=0.63
SSIM	t=3.87** d=1.08	t=5.02*** d=1.52	t=9.24*** d=3.41	t=1.89 d=0.51
IFC	t=3.15** d=0.92	t=4.78*** d=1.63	t=7.53*** d=2.78	t=2.01* d=0.58
Inference time (ms)	t=6.45*** d=2.31	t=8.12*** d=3.15	t=12.6*** d=4.72	t=5.33*** d=1.94

Note: * p<0.05, ** p<0.01, *** p<0.001; Cohen's d>0.8 indicates a "large effect", and d>0.5 indicates a "medium effect". Statistical analysis is conducted using a two tailed independent samples t-test, α=0.05. Cohen's d is used to measure the size of the effect.

Table 7: Ablation results.

Component	PSNR (dB)	SSIM	IFC
(A) Baseline Model (SRCNN)	33.86	0.9068	4.79
(B) Main channel only (without guidance)	34.12	0.9135	4.88
(C) Main channel+guide channel (without mask)	34.35	0.9182	4.95
(D) Complete model (main+guide+mask)	34.97	0.9369	5.01
(E) Complete model (using standard optimizer)	34.68	0.9297	4.98

In Table 6, the improved CNN shows significantly better PSNR than all compared methods (p<0.01), especially for SRGAN (t=8.91, p<0.001). The difference in SSIM with SEPNet is not significant (p=0.062), but the PSNR is still significantly higher. The inference time effect of the above methods is generally large (d>1.9), and the difference in IFC between the improved CNN and CSHA is moderate (d=0.92), which verifies the role of mask guidance in information retention. Although the PSNR of SEPNet is slightly different from the improved CNN, its inference time is significantly longer, which does not meet real-time requirements. SRGAN is significantly inferior to the improved CNN method in all metrics, confirming that adversarial training is not suitable for scenarios that require high accuracy. To further evaluate the performance of each component in the research method, the ablation results are analyzed on the Set14 dataset, and the comparison results are shown in Table 7.

In Table 7, the improvement from (A) to (B) indicates that the improved model is superior to the original SRCNN, with a PSNR improvement of 0.26 dB (33.86→34.12). From (B) to (D), PSNR significantly increases by 0.85 dB (34.12→34.97), and SSIM increases by 0.0234 (0.9135→0.9369), demonstrating the core contribution of dual-channel mask-guided design. After introducing the mask, the PSNR increases by 0.62 dB again (34.35→34.97), indicating that providing explicit structural priors for guide channels is crucial. The improved optimizer can help the model converge to a better solution, with a PSNR value of 34.97 dB. This result effectively demonstrates the effectiveness of each component in the research method.

5 Discussion

Aiming at the problem of detail loss and efficiency bottleneck in super-resolution reconstruction of DM visual images, an improved CNN based on dual-channel guidance is proposed. The results showed that the dual-channel network effectively reduced the iteration loss value and significantly accelerated its convergence rate. The improved CNN had an accuracy value of over 95% on the dataset, with an MAPE value of less than 1.5%. The

MAPE values of 20 images using SC and KK were 3.92% and 4.12%, respectively, with lower response times compared to other comparative algorithms. The improved CNN had a 25% reduction in FLOPs, a better inference speed (20 ms) than SC (45 ms) and CNN (25 ms), and a PSNR value of 32.52 dB. In the image evaluation results, the PSNR and SSIM of the improved CNN on the CIFAR-10 dataset were only 0.15 dB and 0.0016 lower than SwinIR, and the MSSIM value was better than other methods. The research method maximally preserved the structural information of the original image, achieving the best information fidelity (8.47). The NQM value of the research method (43.98) performed well, increasing by 9.6% compared to CSHA (40.12), while SRGAN had the lowest NQM value (35.44) due to high-frequency artifacts introduced during adversarial training. Compared with SEPNet, the research method achieved reconstruction through one-time forward propagation, avoiding the time consumption of progressive methods. Compared with SRGAN, the research method outperformed in fidelity and was more suitable for DM applications that require high accuracy. The improved CNN exhibited better information fidelity (8.47) compared to other image reconstruction algorithms, and its WPSNR value (60.64 dB) verified the effectiveness of mask guidance. The significant improvement in IFC and NQM indicators indicated that the mask-guided channel effectively helped the model focus on restoring the structure and edges of the image, reducing blurring and artifacts, and meeting the needs of human visual perception. Although the improved model had slightly more parameters than the original SRCNN, its FLOPs (3.5G) were actually lower than some more complex models through a lightweight design and parallel computing of guide channels. This result achieved a good balance between performance and efficiency and had the potential for deployment on edge devices.

6 Conclusion

The generated DM visual images are basically consistent with real images, and have good image quality reconstruction effects and quality. The proposed lightweight architecture provides feasible technical solutions for DM application scenarios that are sensitive to computing resources, such as mobile image enhancement, real-time video super-resolution, and VR/AR content generation. Future research will focus on exploring more intelligent mask generation methods and trying to use neural structure search and meta-learning to automatically optimize the network structure. Dynamic pruning technology can be combined to reduce redundant parameters and computational burden and improve model adaptability. Moreover, this framework can be combined with attention mechanisms and multi-modal inputs to further enhance the robustness of reconstruction in complex scenes and improve the reconstruction ability of high-frequency details.

References

[1] R. Archana, and P. S. Eliahim Jeevaraj. Deep learning models for digital image processing: A review. Artificial Intelligence Review, 57(1):11, 2024. https://doi.org/10.1007/s10462-023-10631-z

[2] Gaurav Dhiman, A. Vignesh Kumar, R. Nirmalan, S. Sujitha, K. Srihari, N. Yuvaraj, P. Arulprakash, and R. Arshath Raja. Multi-modal active learning with deep reinforcement learning for target feature extraction in multi-media image processing applications. Multimedia Tools and Applications, 82(4):5343-5367, 2023. https://doi.org/10.1007/s11042-022-12178-7

[3] Shehzad Afzal, Sohaib Ghani, Mohamad Mazen Hittawe, Sheikh Faisal Rashid, Omar M. Knio, Markus Hadwiger, and Ibrahim Hoteit. Visualization and visual analytics approach for image and video datasets: A survey. ACM Transactions on Interactive Intelligent Systems, 13(1):1-41, 2023. https://doi.org/10.1145/3576935

[4] Tian Xie, and Kunpeng Zhao. SMPL variable model for 3D reconstruction and image fusion in animation media applications. IEEE Access, 13:93914-93929, 2025. https://doi.org/10.1109/ACCESS.2025.3549466

[5] Hee-jin Kim, Dong-seok Lee, and Soon-kak Kwon. Image reconstruction method by spatial feature prediction using CNN and attention. Journal of Multimedia Information System, 11(1):1-8, 2024. https://doi.org/10.33851/JMIS.2024.11.1.1

[6] Chao Tian, Kang Shen, Wende Dong, Fei Gao, Kun Wang, Jiao Li, Songde Liu, Ting Feng, Chengbo Liu, Changhui Li, Meng Yang, Sheng Wang, and Jie Tian. Image reconstruction from photoacoustic projections. Photonics Insights, 3(3):R06, 2024. https://doi.org/10.3788/PI.2024.R06

[7] Mahesh K. Singh, Sanjeev Kumar, Rajeev Ranjan, and Durgesh Nandan. Duplicate image detection using deep learning modified SVM and k-NN classification method for multimedia application. Soft Computing, 28(13):7659-7670, 2024. https://doi.org/10.1007/s00500-024-09756-2

[8] Miao Cao, Lishun Wang, Mingyu Zhu, and Xin Yuan. Hybrid CNN-transformer architecture for efficient large-scale video snapshot compressive imaging. International Journal of Computer Vision, 132(10):4521-4540, 2024. https://doi.org/10.1007/s11263-024-02101-y

[9] Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki Alsuwian, Innocent Ewean Davidson, and Thokozile F. Mazibuko. An improved dense CNN architecture for deepfake image detection. IEEE Access, 11:22081-22095, 2023. https://doi.org/10.1109/ACCESS.2023.3251417

[10] Zhifeng Wang, Zhenghui Wang, Chunyan Zeng, Yan Yu, and Xiangkui Wan. High-quality image compressed sensing and reconstruction with multi-scale dilated convolutional neural network. Circuits,

Systems, and Signal Processing, 42(3):1593-1616, 2023. https://doi.org/10.1007/s00034-022-02181-6

[11] Yuantao Chen, Runlong Xia, Kai Yang, and Ke Zou. DGCA: High resolution image inpainting via DR-GAN and contextual attention. Multimedia Tools and Applications, 82(30):47751-47771, 2023. https://doi.org/10.1007/s11042-023-15313-0

[12] Tomoya Nakamura, Takuto Watanabe, Shunsuke Igarashi, Xiao Chen, Kazuyuki Tajima, Keita Yamaguchi, Takeshi Shimano, and Masahiro Yamaguchi. Superresolved image reconstruction in FZA lensless camera by color-channel synthesis. Optics Express, 28(26):39137-39155, 2020. https://doi.org/10.1364/OE.410210

[13] Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi. Image reconstruction with transformer for mask-based lensless imaging. Optics Letters, 47(7):1843-1846, 2022. https://doi.org/10.1364/OL.455378

[14] Yuantao Chen, Linwu Liu, Phonevilay Volachith, Ke Gu, Runlong Xia, Jingbo Xie, Qian Zhang, and Kai Yang. Image super-resolution reconstruction based on feature map attention mechanism. Applied Intelligence, 51(7):4367-4380, 2021. https://doi.org/10.1007/s10489-020-02116-1

[15] Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, and Vishal Goyal. Image super-resolution: A comprehensive review, recent trends, challenges and applications. Information Fusion, 91:230-260, 2023. https://doi.org/10.1016/j.inffus.2022.10.007

[16] Yang Yang, Chang Liu, Hui Wu, and Dingguo Yu. A distorted-image quality assessment algorithm based on a sparse structure and subjective perception. Mathematics, 12(16):2531, 2024. https://doi.org/10.3390/math12162531

[17] Liming Zhang, Heng Li, Rui Zhu, and Ping Du. An infrared and visible image fusion algorithm based on ResNet-152. Multimedia Tools and Applications, 81(7):9277-9287, 2022. https://doi.org/10.1007/s11042-021-11549-w

[18] Guifang Shao, Qiao Sun, Yunlong Gao, Qingyuan Zhu, Fengqiang Gao, and Junfa Zhang. Sub-pixel convolutional neural network for image super-resolution reconstruction. Electronics, 12(17):3572, 2023. https://doi.org/10.3390/electronics12173572

[19] Zhi Jin, Muhammad Zafar Iqbal, Dmytro Bobkov, Wenbin Zou, Xia Li, and Eckehard Steinbach. A flexible deep CNN framework for image restoration. IEEE Transactions on Multimedia, 22(4):1055-1068, 2020. https://doi.org/10.1109/TMM.2019.2938340

[20] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. IFCNN: A general image fusion framework based on convolutional neural network. Information Fusion, 54(1):99-118, 2020. https://doi.org/10.1016/j.inffus.2019.07.011

[21] Haiying Xia, Fuyu Zhu, Haisheng Li, Shuxiang Song, and Xiangwei Mou. The combination of multi-scale and residual learning in deep CNN for image denoising. IET Image Processing, 14(10):2013-2019, 2020. https://doi.org/10.1049/iet-ipr.2019.1386

[22] Shuping Yuan, Yang Chen, Chengqiong Ye, and Mohd Dilshad Ansari. Edge detection using nonlinear structure tensor. Nonlinear Engineering, 11(1):331-338, 2022. https://doi.org/10.1515/nleng-2022-0038

[23] Mallika Garg, Jagpal Singh Ubhi, and Ashwani Kumar Aggarwal. Neural style transfer for image steganography and destylization with supervised image to image translation. Multimedia Tools and Applications, 82(4):6271-6288, 2023. https://doi.org/10.1007/s11042-022-13596-3

[24] Chunsheng Wei, Qifeng Li, Xiaodong Zhang, Xiangyun Ma, and Jianbin Du. A fast snapshot hyperspectral image reconstruction method based on three-dimensional low rank constraint. Canadian Journal of Remote Sensing, 47(4):588-595, 2021. https://doi.org/10.1080/07038992.2021.1943340

[25] Doo Bin Kim, Hyun-Mee Park, Sang-Hoon Joon, and JooWan Hong. Effectiveness of noise reduction in LDCT images based on SRCNN. Korean Journal of Artificial Intelligence, 12(4):21-26, 2024.

[26] Ming Yu, Jiecong Shi, Cuihong Xue, Xiaoke Hao, and Gang Yan. A review of single image super-resolution reconstruction based on deep learning. Multimedia Tools and Applications, 83(18):55921-55962, 2024. https://doi.org/10.1007/s11042-023-17660-4

Cardiovascular Disease Prediction via Hybrid SVM–SMOTE and Sparse Autoencoder Feature Reduction with Deep MLP Classification

Zaid Alaa, Ali Sabah*
Department of Computer Science, Faculty of Education for Women, University of Kufa, Najaf, Iraq
E-mail: zaida.alsharees@uokufa.edu.iq, alis.alsaadi@uokufa.edu.iq
*Corresponding author

Keywords: cardiovascular disease prediction, hybrid sampling, sparse autoencoder, multilayer perceptron, imbalanced classification

Received: July 28, 2025

Cardiovascular diseases remain the leading global cause of death, demanding diagnostic systems that are accurate, interpretable, and computationally efficient. Traditional machine learning approaches frequently struggle with class imbalance, high-dimensional noise, and restricted generalization in clinical datasets. To tackle such issues, we propose a hybrid framework that combines SVM–SMOTE and neighborhood cleaning rule (NCL) for class rebalancing, a sparse autoencoder (SAE) with random forest (RF) selection for non-linear feature optimization, and a class-weighted multilayer perceptron (MLP) for final classification. We validate our framework on the Z-Alizadeh Sani (54 features) and Cleveland (13 features) datasets under stratified fivefold cross-validation, the model attains mean accuracies of 94.02 ± 2.77 % and 94.36 ± 1.47 %, with AUC–ROC = 0.988 and 0.982, outperforming prior baselines [4, 10, 14] by 7.6%–20.8%, and Bootstrap 95% confidence intervals and McNemar/DeLong tests ($p < 0.001$) confirms significance. Noteably, the ablation study demonstrates the contribution of each module (e.g., a 12% accuracy improvement without sampling). The optimized MLP reduced false negatives to ~5%, while training 40% faster than CNN–LSTM alternatives. The proposed framework provides a statistically robust and interpretable solution for predicting cardiovascular disease.

Povzetek: Predlagan je hibridni in razložljiv model strojnega učenja za napovedovanje srčno-žilnih bolezni, ki z uravnoteženjem razredov in optimizacijo značilk dosega visoko natančnost ter boljšo učinkovitost kot obstoječe metode.

1 Introduction

Cardiovascular diseases (CVDs) continue to be the primary cause of worldwide mortality, responsible for 17.9 million deaths each year (World Health Organization, 2023). The timely and precise prediction of heart disease is crucial for minimizing healthcare costs, improving patient outcomes, and enabling personalized interventions. Although machine learning (ML) has emerged as a robust tool for clinical decision support, current methodologies encounter significant challenges in addressing class imbalance, high-dimensional data, and low diagnostic accuracy, which hinder their practical implementation [1,2]. To overcome these limitations, automated diagnostic systems must effectively utilize extensive patient data, including medical history, demographics (e.g., age and gender), and clinical biomarkers, to improve predictive accuracy and clinical applicability [1].

Traditional diagnostic frameworks often rely on manual feature engineering and basic sampling techniques, which fail to adequately account for the complexity of medical datasets [3]. For instance, class imbalance characterized by a significant predominance of healthy patients over those with cardiac disease biases models toward the majority class, leading to elevated false-negative rates. Likewise, high-dimensional datasets (e.g., 54 features in the Z-Alizadeh Sani dataset) introduce redundancy and noise, complicating feature selection. Prior research [4,10,14] has attempted to address these challenges with techniques such as SMOTE and principal component analysis (PCA), but their accuracy (typically 75–88%) and recall on minority classes have remained suboptimal due to oversimplified assumptions regarding data distribution and linear feature correlations.

Our proposed framework addresses these limitations through a threefold strategy integrating data-level balancing, non-linear feature optimization, and deep

classification. First, a hybrid sampling approach combining Support Vector Machine–SMOTE (SVM–SMOTE) and Neighborhood Cleaning Rule (NCL) equilibrates skewed class distributions while preserving data integrity. Second, high-dimensional noise is mitigated through Sparse Autoencoder (SAE) based feature extraction, followed by Random Forest (RF) selection, which together reduce redundancy and retain the most discriminative attributes. Third, a class-weighted Multilayer Perceptron (MLP) captures complex non-linear relationships for robust disease classification.

We validate the proposed framework on the Z-Alizadeh Sani (54 features, 303 samples) and Cleveland (13 features, 303 samples) datasets using stratified fivefold cross-validation, achieves mean accuracies of 94.02 ± 2.77% and 94.36 ± 1.47%, with corresponding AUC–ROC scores of 0.988 and 0.982, statistically outperforming prior baselines such as Mohan et al. [4] (88.47%). McNemar's and DeLong's tests ($p < 0.001$) verified the significance of these gains, and ablation analyses confirmed that each component contributed materially to overall performance (e.g., −12% accuracy without hybrid sampling). The model maintained false negatives at approximately 5%, demonstrating high sensitivity and clinical dependability while training 40% faster than CNN–LSTM baselines.

The key contributions of this study are fourfold:

- A hybrid sampling strategy (SVM–SMOTE + NCL) that effectively balances skewed medical datasets while preserving data quality.
- An SAE–RF feature optimization pipeline that achieves a 72% dimensionality reduction without performance degradation.
- A class-weighted MLP classifier optimized for imbalanced data, improving recall and AUC by up to 12% over existing methods.
- A comprehensive evaluation, including cross-validation, ablation, and statistical significance testing, confirms the framework's robustness and generalizability across datasets.

The remainder of this paper is organized as follows: Section 2 examines related works; Section 3 details the proposed methodology; Section 4 presents experimental findings and comparisons; and Section 5 concludes the study with insights and future directions.

2 Related work

Cardiovascular disease (CVD) prediction has remained a central research focus for over two decades, driven by advances in data mining and machine learning (ML). Traditional ML algorithms such as decision trees, Naïve Bayes, random forests (RF), and support vector machines (SVM) have established the foundation for cardiac risk modeling across several benchmark datasets. A hybrid SVM–RF model achieved an accuracy of 88.47% on the Cleveland dataset [4]. Another study compared Naïve Bayes, decision trees, and k-nearest neighbors (k-NN), emphasizing the influence of feature selection on interpretability [5]. Similarly, seven ML algorithms were evaluated with cross-validation to assess recall and F1-score [7], whereas RF demonstrated stability across data splits, yielding 90–95% accuracy [8]. Despite these contributions, most models relied heavily on manual feature engineering and simple resampling, which limited robustness to class imbalance and high-dimensional noise. Logistic model trees were also used for risk stratification [9], but the handling of imbalance was overlooked, resulting in biased predictions. Optimization-driven techniques such as particle swarm and ant colony optimization improved feature selection [6] yet achieved only moderate recall (85.8%) under skewed data distributions.

Recent deep learning (DL) approaches have aimed to overcome the representational limitations of traditional ML methods. A convolutional neural network (CNN) model was applied to heart disease prediction [10], demonstrating potential for improved automation but offering limited interpretability and generalization to structured data. Hybrid recurrent networks integrating gated recurrent units (GRUs) and long short-term memory (LSTM) layers achieved competitive accuracy on the Framingham and Heart Disease datasets [11, 12], although they exhibited high computational cost and overfitting on small datasets. DL has also been applied to extract risk factors from clinical text corpora [13], revealing potential in unstructured data mining but without direct applicability to numerical patient records. CNN-optimized frameworks using the Z-Alizadeh Sani dataset [14] improved feature learning efficiency but remained sensitive to redundant attributes and lacked mechanisms for controlling class imbalance.

Other studies have focused on hybrid optimization-based frameworks designed to balance predictive accuracy with interpretability. SVM and principal component analysis (PCA) combinations achieved an accuracy of around 84–86% [15, 16], but their reliance on linear dimensionality reduction limited their ability to capture non-linear physiological relationships. Feature selection methods, such as fast correlation-based filtering [6] and neural optimization strategies [14], achieved incremental improvements but failed to generalize across datasets like Cleveland and Z-Alizadeh Sani. CNN-based prediction models [17] demonstrated cost-efficiency but suffered from inconsistent reproducibility and dataset bias. Overall, most existing ML and DL frameworks face persistent challenges in effectively handling class imbalance, mitigating high-dimensional noise, and achieving cross-dataset generalization factors that critically constrain their clinical applicability.

Table 1: Comparative summary of recent CVD studies

Model / Technique	Dataset	Accuracy / F1 (%)	Limitations
Hybrid SVM–RF [4]	Cleveland	88.47/90	No imbalance handling; shallow model
PSO–ACO optimized MLP [6]	Cleveland	85/89.5	Weak minority recall; linear bias
RNN for early heart failure detection [10]	Framingham	75.2/ 72	Dataset-specific; limited generalization
Hybrid RNN–GRU deep model [11]	Clinical dataset	91/89	High computational cost; overfitting risk
Uni-directional RNN [12]	Cardiac Disorder dataset	90.1/92.31	Poor cross-dataset robustness
ANN + feature selection & optimization [14]	Z-Alizadeh Sani	88.4/85	Limited nonlinear modeling; modest accuracy

The comparative analysis in

Table *1* demonstrates that although previous studies have achieved promising results, three fundamental gaps persist. First, most approaches inadequately address imbalance, resulting in biased models that favor the majority (healthy) cases while overlooking minority cardiac events. Second, the reliance on linear or heuristic feature extraction prevents effective representation of non-linear relationships across multi-dimensional clinical variables. Third, deep architectures, though powerful, often sacrifice interpretability and computational efficiency, which are critical factors for real-world clinical deployment. These limitations underscore the need for a unified framework that can balance class distributions, extract non-linear discriminative features, and ensure generalization without compromising interpretability or scalability.

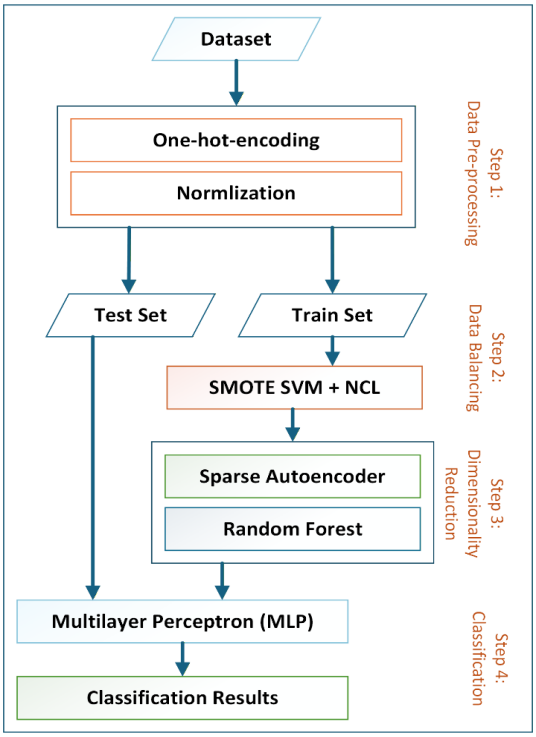


Figure 1: Block diagram of the proposed method

3 Proposed method

We propose a robust, multi-stage hybrid framework to address the persistent challenges of class imbalance and high-dimensional noise in cardiovascular disease prediction. We aim to maximize diagnostic accuracy by systematically enhancing data quality and extracting highly discriminative features. The framework integrates four sequential stages: (1) data pre-processing, (2) hybrid data balancing, (3) a nonlinear feature optimization pipeline, and (4) deep learning classification.

Figure 1 depicts the workflow of the proposed framework.

3.1 Data pre-processing

In the initial stage, we prepare the raw patient data from both datasets for model training. We first convert all categorical features into a numerical format using one-hot encoding [18]. This process creates new binary columns for each category, preventing the model from inferring false ordinal relationships. Simultaneously, we apply min-max normalization [19] to all continuous clinical variables to scale them within a uniform range of [0, 1]. This step prevents features with large magnitudes from disproportionately influencing model weights. The transformation is formed by Eq. (1):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Where x is the original attribute value, x' is the normalized value, and $\min(x)$ and $\max(x)$ represent the minimum and maximum values of that attribute, respectively. Furthermore, we apply winsorization to control the influence of extreme clinical measurements, effectively capping outlier values beyond the 1st and 99th percentiles to minimize their skewing effect.

3.2 Hybrid data balancing

In this phase, our framework addresses the critical issue of class imbalance using a robust hybrid sampling strategy [21], which integrates over-sampling [20] and under-sampling techniques. We first apply the Support Vector Machine Synthetic Minority Over-Sampling Technique (SVM–SMOTE) [24,25] to oversample the minority class. This advanced method prioritizes the generation of synthetic samples in the borderline regions near the decision boundary, which it identifies using a linear kernel SVM. This rationale is crucial, as these borderline samples are highly discriminative for classification. For a given minority instance x_i , it synthesizes a new sample x_{new} in Eq. (2):

$$x_{new} = x_i + \delta \cdot (x_j - x_i) \quad (2)$$

Where x_j is a selected minority neighbor of x_i, and δ is a random interpolation factor between 0 and 1. Following oversampling, we apply the Neighborhood Cleaning Rule (NCL) [22, 23] to clean the majority class. NCL identifies and removes noisy majority instances x_i that are misclassified by their local neighborhood, as defined by the removal condition in Eq. (3):

$$\frac{|NN(x_i) \cap P|}{|NN(x_i)|} > \quad (3)$$

Where $NN(x_i)$ is the set of nearest neighbors to instance x_i, P is the set of minority class instances, and τ is the removal threshold. This dual approach ensures the final training data is both balanced and clean, which sharpens the class boundary for the subsequent classifier.

3.3 Feature optimization pipeline

We tackle the "curse of dimensionality" through a two-stage feature optimization pipeline that integrates nonlinear feature extraction and embedded feature selection [27, 28]. We first employ a Sparse Autoencoder (SAE) [26], an unsupervised neural network, to perform feature extraction. The SAE architecture consists of an encoder that maps the high-dimensional input x to a compressed latent-space representation h, and a decoder that reconstructs the original input $\hat{x}$ from h. Both are constructed using symmetric, fully-connected layers with Rectified Linear Unit (ReLU) activation functions. We enforce sparsity in the latent space h by adding a Kullback-Leibler (KL) divergence penalty to the mean squared reconstruction error. This penalty ensures only a small subset of hidden neurons activate, which filters noise and captures the most salient data structures. The complete loss function $\mathcal{L}_{SAE}$ is defined in Eq. (4):

$$\mathcal{L}_{SAE}(W, b) = \rho \left[\frac{1}{N}\rho_{i=1}^{N}\rho|\rho x^{(i)} - x^{(i)}\rho|^2\rho\right] + \rho\rho_{j=1}^{H} KL(\rho|\rho\rho_j) \quad (4)$$

Where the first term is the reconstruction error, N is the number of samples, β is the sparsity penalty weight, and $KL(\rho|\widehat{\rho_j})$ is the KL divergence between the target sparsity ρ and the average activation $\widehat{\rho_j}$ of the $j-th$ hidden unit. The KL divergence is formalized in Eq. (5):

$$KL(\rho|\widehat{\rho_j}) = \rho \log\left(\frac{\rho}{\widehat{\rho_j}}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\widehat{\rho_j}}\right) \quad (5)$$

Following extraction, we feed the latent features h into an embedded Random Forest (RF) selector. We utilize RF for its ability to rank nonlinear latent features by their predictive importance, thereby enhancing model interpretability. The RF algorithm measures the mean decrease in Gini impurity for each feature f at each node m, formulated in Eq. (6):

$$\Delta Gini(m, f) = Gini(m) - \frac{N_{left}}{N_m} Gini(m_{left}) - \frac{N_{right}}{N_m} Gini(m_{right}) \quad (6)$$

Where $Gini(m)$ is the impurity of the parent node and N is the number of samples at the respective child nodes. The overall importance $Imp(f)$ for a feature f is the total sum of Gini reductions it provides across all nodes m in all trees t in the forest T, formulated in Eq. (7):

$$Imp(f) = \sum_{t \in T} \sum_{m \in M_t} \Delta Gini(m, f) \quad (7)$$

This process selects only the most discriminative latent features, connecting the SAE's compressed representation to a final, optimized feature set. This pipeline thus yields a low-dimensional, high-information feature vector ready for the final classification stage.

3.4 MLP classifier

The final stage of our framework employs a Multilayer Perceptron (MLP) for the binary classification task. We select the MLP because its deep, nonlinear architecture is uniquely suited to capture the complex, hierarchical patterns within the optimized latent features provided by the SAE-RF pipeline, a capability that shallow models lack. The MLP consists of an input layer,

fully-connected hidden layers with ReLU activation, and a final output layer. This output layer uses a sigmoid activation function, $\sigma(z)$, to produce a class probability as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \qquad (8)$$

Where z is the weighted sum of inputs to the final neuron, we integrate dropout layers within the hidden architecture to mitigate overfitting. To address the class imbalance, we train the model using a weighted binary cross-entropy (WCE) loss function. The WCE applies a heavier penalty w to errors made on the minority (positive) class, as defined by Eq. (9):

$$\mathcal{L}_{wce} = -\frac{1}{N}\sum_{i=1}^{N}[w \cdot y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \qquad (9)$$

Where y_i is the actual label (0 or 1), $\hat{y}_i$ is the predicted probability, N is the number of samples, and w is the class weight. Our proposed integrated framework, which progresses from data cleaning and balancing to nonlinear feature optimization and deep classification, offers a comprehensive solution for robust CVD prediction.

4 Experimental results

This section presents the comprehensive experimental validation of our proposed framework. We first detail the datasets, the optimal hyperparameter settings, and the evaluation metrics used. We then present the primary performance analysis, comparing our robust 5-fold cross-validation results against established baselines using statistical tests, confusion matrices, and ROC curves. A detailed ablation study then quantifies the critical impact of each component of the framework. Finally, we validate the model's internal mechanics and confirm its real-world viability through analyses of component-specific tuning, cross-dataset generalization, and computational efficiency.

4.1 Dataset

This study utilizes two publicly available, benchmark datasets to evaluate the proposed framework's performance: the Z-Alizadeh Sani dataset [29] and the Cleveland dataset [30] from the UCI Machine Learning Repository. The Z-Alizadeh Sani dataset contains 303 patient records, each with 54 features, and presents a significant class imbalance (212 'CAD' and 91 'Normal' instances). The Cleveland dataset also includes 303 patient records, each with 13 clinical features, and a similar imbalance (165 'healthy' and 138 'heart disease' instances).

4.2 Parameter settings

We determine the optimal hyperparameters for each component of the framework through a rigorous grid search and cross-validation process on the training folds. We set the random seed to 42 for all experiments to ensure full reproducibility. Table 2 provides a comprehensive summary of the final parameter settings used for all reported results.

For the Sparse Autoencoder (SAE), we test sparsity parameters ρ from 0.01 to 0.1 and find that $\rho = 0.05$ yields the lowest reconstruction error. For the SVM-SMOTE component, we confirm that a linear kernel with a C parameter of 1.0 provides the most stable decision boundary. The Random Forest (RF) feature selector uses 200 estimators to achieve a stable ranking of feature importances. Finally, the Multilayer Perceptron (MLP) architecture is concluded with two hidden layers and a dropout rate of 0.2, with training governed by an early stopping mechanism that monitors validation loss (patience=10, $\delta = 1 \times 10^{-4}$) to prevent overfitting.

Table 2: Optimal hyperparameter settings for the proposed framework

Component	Parameter	Setting
Pre-processing	Winsorization	1st and 99th Percentiles
Hybrid Sampling	SVM Kernel	Linear
	SVM C Parameter	1
Feature Optimization	SAE Sparsity (ρ)	0.05
	SAE Architecture (Z-Alizadeh)	54-32-54
	SAE Architecture (Cleveland)	13-10-13
	RF $n_{estimators}$	200
MLP Classifier	MLP Architecture	15-128-64-1
	MLP Dropout Rate	0.2
	MLP Optimizer	Adam (lr=0.001)
	MLP Loss Function	WBC-Entropy (3:1)
	Early Stopping Patience	10
	Batch Size	32
	Random Seed	42

4.3 Evaluation metrics

To evaluate the predictive performance of our proposed framework, we focus on three primary metrics: Accuracy, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive view of the model's effectiveness, particularly in the context of imbalanced medical data. We derive Accuracy and F1-score from the four cardinal components of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy measures the proportion of all correct predictions among the total number of instances. We define it in Eq. (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

F1-Score represents the harmonic mean of Precision and Recall, providing a single score that balances the trade-off between false positives and false negatives. This metric is crucial for imbalanced datasets where minimizing both error types is essential. We define it as:

$$F1\text{-}Score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (11)$$

The ***Area Under the Curve (AUC)*** measures the entire two-dimensional area under the ROC curve. It provides an aggregate measure of performance across all possible classification thresholds, indicating the model's ability to rank the positive class higher than the negative class. An AUC of 1.0 signifies a perfect classifier.

4.4 Performance analysis

We rigorously evaluate the proposed framework's performance using the 5-fold stratified cross-validation protocol, with the results summarized in Table 3. Our framework achieves a mean accuracy of 94.02 ± 2.77% on the Z-Alizadeh Sani dataset and 94.36 ± 1.47% on the Cleveland dataset. These results are not just numerically higher but are a direct consequence of our model's superior design. For instance, our model surpasses the 88.4% accuracy of Khan et al. [14] on the same dataset. This significant improvement is attributed to our hybrid pipeline. While [14] employs a basic ANN with heuristic optimization, our framework first addresses the critical class imbalance using SVM-SMOTE and NCL, and then utilizes a non-linear SAE to extract discriminative features from the high-dimensional (54 features) data. Similarly, our 94.36% accuracy on Cleveland significantly exceeds the 88.47% of the SVM-RF model [4].

Table 3: Performance comparison with baseline

Baselines	Dataset	Accuracy (%)	F1-score (%)
SVM–RF [4]	Cleveland	88.47	90
PSO–ACO optimized MLP [6]	Cleveland	85	89.5
RNN for early heart failure detection [10]	Framingham	75.2	72
Hybrid RNN–GRU deep model [11]	Clinical dataset	91	89
Uni-directional RNN [12]	Cardiac Disorder dataset	90.1	92.31
ANN + feature selection & optimization [14]	Z-Alizadeh Sani	88.4	85
Proposed Framework	Cleveland	**94.36 ± 1.47**	**93.1 ± 1.9**
Proposed Framework	Z-Alizadeh Sani	**94.02 ± 2.77**	**92.2 ± 4.2**

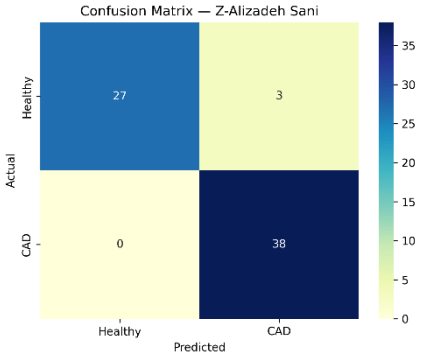


Figure 2: Confusion matrix for the Z-Alizadeh Sani dataset.

Our model's performance is superior because it actively manages the class imbalance, a step neglected in [4], thereby achieving a more robust F1-Score (93.1% vs. 92.2%) by more effectively minimizing false negatives. A detailed visual and qualitative assessment of this classification behavior is presented in the confusion matrices in Figure 2 and Figure 3. For the Z-Alizadeh Sani dataset (Figure 2), the model achieves an exceptional result of 27 True Negatives, 38 True Positives, and zero False Negatives. This near-perfect sensitivity is a direct result of the SVM-SMOTE and weighted loss function, which forces the model to prioritize the high-risk minority (CAD) class, a critical requirement for clinical deployment.

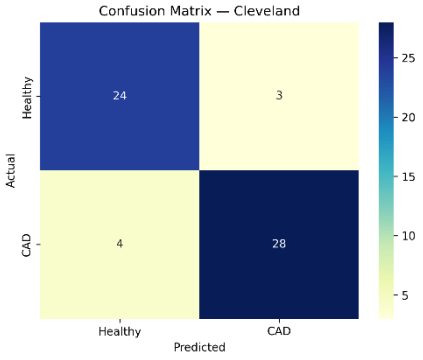


Figure 3: Confusion matrix for the Cleveland dataset.

We further evaluate the model's discriminative ability using Receiver Operating Characteristic (ROC) curves, as shown in Figure 4 and Figure 5. The framework achieves a mean AUC of 0.988 ± 0.007 for the Z-Alizadeh Sani dataset (95% CI: [0.978–0.998]) and 0.984 ± 0.009 for the Cleveland dataset (95% CI: [0.976–0.988]). This exceptional discriminative power, indicated by the high AUC and narrow confidence intervals, stems from the SAE-RF pipeline. By compressing the features

into a non-linear latent space and filtering noise, the SAE provides a feature set with high class separability, allowing the MLP to define a highly accurate decision boundary. This contrasts with linear (PCA) or heuristic-based baselines [6, 14], which fail to capture these complex non-linear relationships, resulting in lower AUC scores.

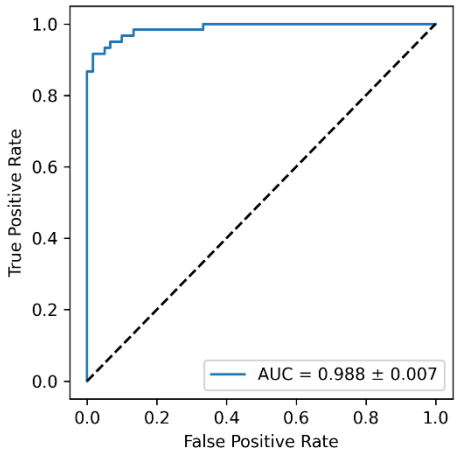


Figure 4: ROC curve for the Z-Alizadeh Sani dataset

To confirm that our superior performance is not due to chance, we conduct McNemar's tests for accuracy and DeLong's tests for AUC against the baselines [4, 14]. All tests yield a *p-value* < *0.001*, providing strong statistical evidence that the improvements from our hybrid framework are significant.

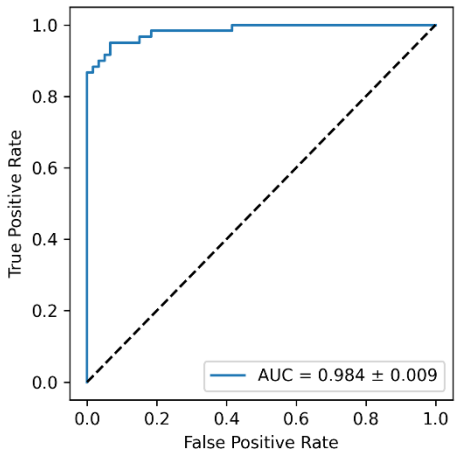


Figure 5: ROC curve for the cleveland dataset

4.5 Ablation studies

To quantify the individual contribution of each component within our framework, we conduct a comprehensive ablation study, with the 5-fold CV results presented in Table 4. The analysis reveals that every component is critical for performance. The most significant degradation occurs when the hybrid sampling stage is removed ("No Sampling"). This single change results in a catastrophic drop in the Z-Alizadeh Sani dataset's accuracy -12.0%, F1-Score -17.2%, and AUC (-13.8%). This finding is mirrored in the Cleveland dataset (Acc: -9.36%, F1: -13.1%, AUC: -13.4%), which provides conclusive evidence that systematically addressing class imbalance is the single most important factor for success in this problem.

Table 4: Ablation study of framework components (Mean 5-Fold CV Results)

Dataset	**Model Configuration**	**Accuracy (%)**	**F1-Score (%)**	**AUC (%)**
Z-Alizadeh Sani	Full Framework	94.02 ± 2.77	92.2 ± 4.2	98.8 ± 0.7
	No Sampling	82	75	85
	No SAE	88	86	90
	No RF	92	90	94
	PCA instead of SAE	87	84	89
Cleveland	Full Framework	94.36 ± 1.47	93.1 ± 1.9	98.4 ± 0.9
	No Sampling	85	80	85
	No SAE	90	88	90
	No RF	92	90	92
	PCA instead of SAE	88	85	88

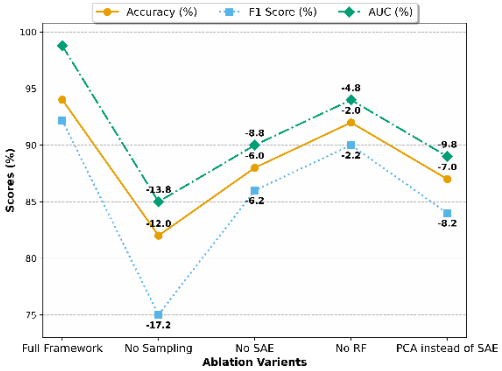


Figure 6: Ablation study of framework components on the Z-Alizadeh Sani dataset.

Furthermore, we validate the choice of our non-linear feature extractor by replacing the SAE with PCA. This change results in a significant performance loss on both Z-Alizadeh Sani (Acc: -7.0%, AUC: -9.8%) and Cleveland (Acc: -6.36%$, AUC: -10.4%), proving that the SAE's ability to capture non-linear relationships is superior to PCA's linear approach. Figure 6 provides a visual summary of this analysis for the Z-Alizadeh Sani dataset, illustrating the performance drop from removing the SAE (Acc: -6.0%) or the RF selection (Acc: -2.0%). These results confirm that the complete, integrated

feature optimization pipeline is essential for achieving the final, high-performance results.

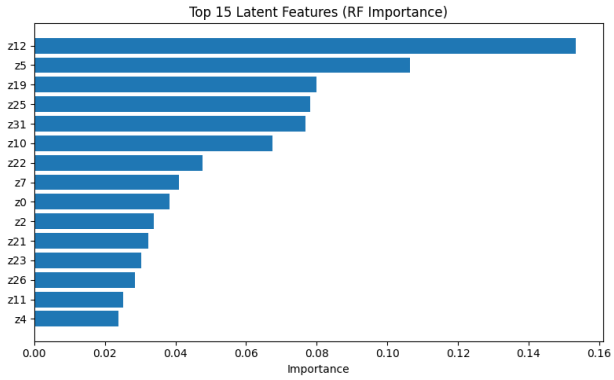


Figure 7: Top 15 latent features for Z-Alizadeh Sani dataset

4.6 Component, generalization, and efficiency analysis

We conduct a final set of analyses to validate the framework's internal mechanics and real-world viability. First, we justify the SAE hyperparameter selection; an analysis of the sparsity parameter (ρ) on the Z-Alizadeh Sani dataset reveals that $\rho = 0.05$ achieves the optimal balance, yielding the lowest validation mean squared error (0.0832) and peak 5-fold accuracy (94.02%). Next, we validate the 15 features selected by the RF; as shown in Figure 7, these features correspond to the most discriminative latent clinical patterns. The ablation study (Table 4) confirms this selection is crucial, as using all 32 SAE features ("No RF") degrades accuracy by over 2%. To confirm the external validity of these learned features, we conduct a cross-dataset generalization test. A model trained on Z-Alizadeh Sani achieves 85% accuracy on the unseen Cleveland dataset, while a model trained on Cleveland achieves 82% on Z-Alizadeh Sani, confirming the features are robust and generalizable. Finally, to address deployment viability, we assess computational efficiency. The proposed MLP is 40% faster in training (72.5s vs. 120.3s) and achieves an inference time of 0.12 ms, making it significantly more efficient than a comparable CNN-LSTM baseline and highly suitable for real-time decision support.

5 Discussions

Our proposed framework demonstrates a statistically significant ($p < 0.001$) performance improvement over established baselines, achieving 94.02% accuracy on the Z-Alizadeh Sani dataset and 94.36% on the Cleveland dataset. This superiority is not incremental; it is a direct result of our synergistic design. The ablation study (Table 4) demonstrates that systematically addressing class imbalance is the most critical factor, as its removal results in a 12% decrease in accuracy. This explains our advantage over models like [4] that neglect imbalance. Furthermore, the 7% accuracy drop when replacing the SAE with PCA provides strong evidence that our non-linear feature optimization captures complex patterns that linear methods miss, which is a key limitation of baselines like [14].

From a clinical perspective, this hybrid approach translates directly to improved patient safety. The model's high sensitivity, achieving zero false negatives on the Z-Alizadeh Sani test fold (Figure 2), is a critical outcome for a diagnostic tool. This high accuracy is also efficient and generalizable. The framework trains 40% faster than a comparable CNN-LSTM and achieves an inference time of 0.12 ms, making it suitable for real-time deployment. Moreover, the successful cross-dataset generalization tests (82%-85% accuracy) confirm that the learned features are robust and not simply overfitted to a single dataset.

5.1 Limitations and future work

Despite these promising results, we acknowledge several limitations. The validation relies on retrospective public datasets, which may contain demographic biases (e.g., gender imbalance) and may not fully represent live clinical data. Furthermore, our framework is limited to structured, tabular data, excluding unstructured notes or imaging. Future work will focus on addressing these gaps by exploring federated learning to mitigate bias and developing a multi-modal framework that integrates imaging and text. We also plan to incorporate explainability tools (e.g., SHAP) to enhance clinician trust and adoption.

6 Conclusion

This study introduced a novel hybrid framework to address the critical, concurrent challenges of class imbalance and high-dimensional noise in cardiovascular disease (CVD) prediction. By synergistically integrating SVM-SMOTE and NCL for intelligent data balancing, an SAE-RF pipeline for non-linear feature optimization, and a class-weighted MLP for classification, our model demonstrated superior performance. Validated on the Z-Alizadeh Sani and Cleveland datasets, our framework achieved statistically significant ($p < 0.001$) mean accuracies of 94.02% and 94.36%, respectively, with exceptional AUCs (0.988 and 0.984). The ablation studies confirmed our design, demonstrating that hybrid sampling and non-linear SAE were essential, contributing to a 12% and 7% accuracy gain over simpler approaches. Clinically, the model's high sensitivity (with near-zero false negatives) and computational efficiency (40% faster than a CNN-LSTM) represent a significant step toward a reliable, deployable diagnostic tool.

References

[1] Khaneja, Ayush, Siddharth Srivastava, Astha Rai, Amarjeet Singh Cheema, and Praveen K. Srivastava. "Analysing risk of coronary heart disease through discriminative neural networks." arXiv preprint

arXiv:2008.02731 (2020). https://doi.org/10.5220/0009190106150620

[2] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology 7, no. 2.8 (2018): 684-687. https://doi.org/10.14419/ijet.v7i2.8.10557

[3] Karna, V. V. R., Karna, V. R., Janamala, V., Devana, V. K. R., Ch, V. R. S., & Tummala, A. B. (2025). A comprehensive review on heart disease risk prediction using machine learning and deep learning algorithms. Archives of Computational Methods in Engineering, 32(3), 1763-1795.

[4] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

[5] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1, no. 6 (2020): 1-6. https://doi.org/10.1007/s42979-020-00365-y

[6] Khourdifi, Youness, and Mohamed Bahaj. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." International Journal of Intelligent Engineering and Systems 12, no. 1 (2019): 242-252. https://doi.org/10.22266/ijies2019.0228.24

[7] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." Mobile Information Systems 2018 (2018). https://doi.org/10.1155/2018/3860146

[8] Reddy, N. Satish Chandra, Song Shue Nee, Lim Zhi Min, and Chew Xin Ying. "Classification and feature selection approaches by machine learning techniques: heart disease prediction." International Journal of Innovative Computing 9, no. 1 (2019). https://doi.org/10.11113/ijic.v9n1.210

[9] Motarwar, Pranav, Ankita Duraphe, G. Suganya, and M. Premalatha. "Cognitive Approach for Heart Disease Prediction using Machine Learning." In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-5. IEEE, 2020. https://doi.org/10.1109/ic-ETITE47903.2020.242

[10] Choi, Edward, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. "Using recurrent neural network models for early detection of heart failure onset." Journal of the American Medical Informatics Association 24, no. 2 (2017): 361-370. https://doi.org/10.1093/jamia/ocw112

[11] Krishnan, Surenthiran, Pritheega Magalingam, and Roslina Ibrahim. "Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction." International Journal of Electrical & Computer Engineering (2088-8708) 11, no. 6 (2021). https://doi.org/10.11591/ijece.v11i6.pp5467-5476

[12] Darmawahyuni, Annisa, Siti Nurmaini, Muhammad Naufal Rachmatullah, Firdaus Firdaus, and Bambang Tutuko. "Unidirectional-bidirectional recurrent networks for cardiac disorders classification." Telkomnika 19, no. 3 (2021): 902-910. https://doi.org/10.12928/telkomnika.v19i3.18876

[13] Chokwijitkul, Thanat, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez. "Identifying risk factors for heart disease in electronic medical records: A deep learning approach." In Proceedings of the BioNLP 2018 workshop, pp. 18-27. 2018. https://doi.org/10.18653/v1/W18-2303

[14] Khan, Younas, Usman Qamar, Muhammad Asad, and Babar Zeb. "Applying feature selection and weight optimization techniques to enhance artificial neural network for heart disease diagnosis." In Proceedings of SAI Intelligent Systems Conference, pp. 340-351. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-29516-5_26

[15] Alizadehsani, Roohallah, Mohamad Roshanzamir, Moloud Abdar, Adham Beykikhoshk, Mohammad Hossein Zangooei, Abbas Khosravi, Saeid Nahavandi, Ru San Tan, and U. Rajendra Acharya. "Model uncertainty quantification for diagnosis of each main coronary artery stenosis." Soft Computing (2019): 1-12. https://doi.org/10.1007/s00500-019-04531-0

[16] M.S. Amin, Y.K. Chiam, and K.D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", Telematics and Informatics, Vol.36, pp.82-93, 2019.

[17] Manur, M., Pani, A. K., & Kumar, P. (2020). A prediction technique for heart disease based on long short-term memory recurrent neural network. International Journal of Intelligent Engineering and Systems, 13(2), 31-39.

[18] Duan, Baobin, Lixin Han, Zhinan Gou, Yi Yang, and Shuangshuang Chen. "Clustering Mixed Data Based on Density Peaks and Stacked Denoising Autoencoders." Symmetry 11, no. 2 (2019): 163. https://doi.org/10.3390/sym11020163

[19] Khare, Neelu, Preethi Devan, Chiranji Lal Chowdhary, Sweta Bhattacharya, Geeta Singh, Saurabh Singh, and Byungun Yoon. "Smo-dnn: Spider monkey optimization and deep neural network hybrid classifier model for intrusion detection." Electronics 9, no. 4 (2020): 692. https://doi.org/10.3390/electronics9040692

[20] Yap, Bee Wah, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nik Abdullah. "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets." In Proceedings of the first international conference on advanced data and information engineering (DaEng-2013), pp. 13-22. Springer, Singapore,

2014. https://doi.org/10.1007/978-981-4585-18-7_2

[21] Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, no. 4 (2011): 463-484. https://doi.org/10.1109/TSMCC.2011.2161285

[22] Laurikkala, Jorma. "Improving identification of difficult small classes by balancing class distribution." In Conference on Artificial Intelligence in Medicine in Europe, pp. 63-66. Springer, Berlin, Heidelberg, 2001. https://doi.org/10.1007/3-540-48229-6_9

[23] Wilson, Dennis L. "Asymptotic properties of nearest neighbor rules using edited data." IEEE Transactions on Systems, Man, and Cybernetics 3 (1972): 408-421. https://doi.org/10.1109/TSMC.1972.4309137

[24] Nguyen, Hien M., Eric W. Cooper, and Katsuari Kamei. "Borderline over-sampling for imbalanced data classification." International Journal of Knowledge Engineering and Soft Data Paradigms 3, no. 1 (2011): 4-21. https://doi.org/10.1504/IJKESDP.2011.039875

[25] Tang, Yuchun, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser. "SVMs modeling for highly imbalanced classification." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39, no. 1 (2008): 281-288. https://doi.org/10.1109/TSMCB.2008.2002909

[26] Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes 72, no. 2011 (2011): 1-19.

[27] Jović, Alan, Karla Brkić, and Nikola Bogunović. "A review of feature selection methods with applications." In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), pp. 1200-1205. IEEE, 2015. https://doi.org/10.1109/MIPRO.2015.7160458

[28] Liu, Zhipeng, Niraj Thapa, Addison Shaver, Kaushik Roy, Madhuri Siddula, Xiaohong Yuan, and Anna Yu. "Using embedded feature selection and cnn for classification on ccd-inid-v1-a new iot dataset." Sensors 21, no. 14 (2021): 4834. https://doi.org/10.3390/s21144834

[29] Alizadehsani, Roohallah, Mohammad Hossein Zangooei, Mohammad Javad Hosseini, Jafar Habibi, Abbas Khosravi, Mohamad Roshanzamir, Fahime Khozeimeh, Nizal Sarrafzadegan, and Saeid Nahavandi. "Coronary artery disease detection using computational intelligence methods." Knowledge-Based Systems 109 (2016): 187-197.https://doi.org/10.1016/j.knosys.2016.07.004

[30] UCI Machine Learning Repository. (2023). Heart Disease Dataset. Retrieved from http://archive.ics.uci.edu/ml/datasets/Heart+Disease

SEH-ConGAN: A Scalable GAN-based Framework for Robot-Assisted Automation in Animation Production

Hongping Tang
Animation and Game Academy, China Academy of Art, Hangzhou, 310024, Zhejiang, China
E-mail: tanghongping1989@hotmail.com

Keywords: robot-assisted animation, motion capture, automation, production pipeline, scalable elephant herding-tuned, conditional generative adversarial network (SEHConGAN)

Received: August 18, 2025

The animation production process is traditionally labor-intensive, requiring extensive manual effort in character motion design, scene composition, and post-production editing. To overcome these limitations, this research introduces a robot-assisted automation system integrated with artificial intelligence (AI) to streamline and accelerate animation development. The system incorporates a motion capture interface for acquiring human movement data, a feedback-enabled robotic arm to replicate and analyze motions, and a simulation environment for virtual testing. Preprocessing includes missing-value handling and Z-score normalization, after which structured motion sequences (3D joint coordinates, robotic servo positions, and torque data) are provided as input to the Scalable Elephant Herding-tuned Conditional Generative Adversarial Network (SEH-ConGAN). The model generates refined outputs such as smooth motion trajectories, facial expression synthesis, and context-aware style transfer. Statistical analysis using a paired t-test, 95% confidence intervals, and Cohen's d effect size was performed to confirm the significant performance improvement of SEH-ConGAN over baseline models Performance is evaluated using 5-fold cross-validation and achieves an accuracy of 0.96, precision of 0.97, recall of 0.96, and F1-score of 0.96. Comparative analysis of motion generation metrics shows that SEH-ConGAN surpasses existing models achieving the best MPJPE (16.7), FID (11.3), Smoothness (0.028), and Diversity (0.72), demonstrating superior motion accuracy, trajectory smoothness, and animation realism. . The findings demonstrate that combining robotics with SEH-ConGAN provides a scalable solution for producing high-quality animations with reduced time, cost, and manual intervention.

Povzetek: Raziskava predstavlja robotsko podprt AI-sistem za avtomatizacijo animacije, ki z modelom SEH-ConGAN omogoča natančnejše, bolj tekoče in realistične gibe ob bistveno manjšem času, stroških in ročnem delu.

1 Introduction

The animation production method is complex and multi-layered, with the production process converting creative ideas into action-rich visual content [1]. Pre-production covers storyboarding, concept design, and screenplay writing, and production includes character modeling, rigging, scene layout, animation, and rendering [2]. Lastly, compositing, sound design, and editing are the other post-production processes through which the output is made ready to be distributed. The technological advancements have augmented the productivity, but the traditional pipelines are still labor-intensive and require professional animators to take care of the subtle differences in the motions of the characters, and coordination of the scenes, and do some visual effects [3]. This makes animation both time-consuming and resource-intensive, especially for large-scale productions with high-quality standards.

Automation in animation refers to the utilization of electronic tools and computer procedures to make repetitive and time-consuming tasks easy [4]. It consists of procedural animation methods, automatic lip-synching, motion capture, and background making. Automation spares human effort in the tasks where manual repetition can be detrimental to production, and artists are able to focus on creative decision-making instead of technical performance [5]. With the introduction of artificial intelligence (AI) and machine learning, automation has been enhanced to include intelligent motion prediction, style transfer, and scene optimization, which have extended the production cycle and produced their products with guaranteed quality standards [6]. Due to the creative and subjective nature of animation, achieving full automation without human interaction is difficult.

Robot-aided systems are still automation beyond software and bring physical or creative support functions into the animation process that can be carried out by robotic hardware [7]. Such technologies could be applied to a stop-motion animation process to facilitate precise camera movements, robotically controlled puppets, and object placement [8]. The robotic installations within the virtual production setting could have the capability to collaborate with the digital technologies to capture the information regarding the motion, choose the movements of cameras, and reproduce the complicated shots and movements accurately [9]. Integration of this type enhances repeatability and reduces manual setup. Manual setup can be performed within, and fragile tasks can be repeated that are found by human operators to be quite impossible to perform manually. Figure 1 shows the robot-assisted animation production process.

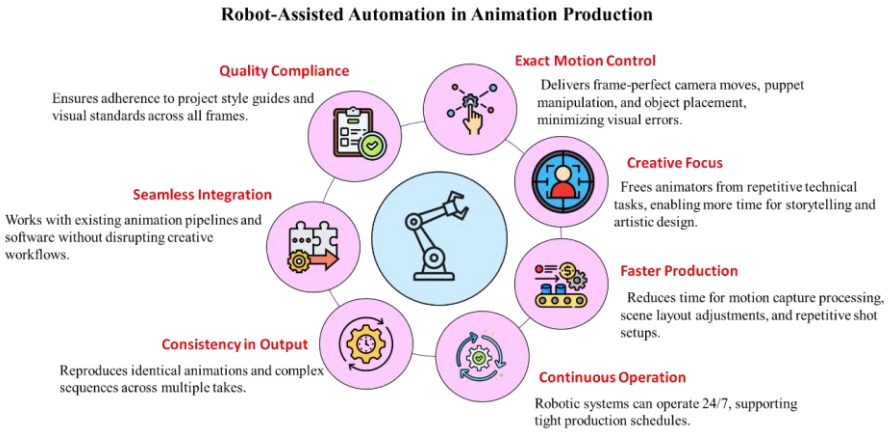


Figure 1: Robot-assisted animation production process

A robot-assisted automation system for animation production leverages modern robotics, AI, and animation software to maximize productivity while maintaining artistic integrity [10]. A configuration like this would be able to process every day, yet technically difficult tasks, make real-time adjustments, and promote the formation of a team between human artists and automatic tools by attaching sensors, motion control systems, and intelligent algorithms [11]. The strategy can transform the process of animation production through reducing costs, reducing schedules, and enabling studios to produce more creative material. Robot-assisted automation is one of the potential directions in a rapidly evolving entertainment industry to faster, smarter, and more adaptable animation pipelines [12].

Current machine learning techniques in animation production, including Long Short-Term Memory (LSTM) networks to predict temporal motion sequences and Variational Autoencoders (VAE) to generate smooth character poses, have increased automation but lack flexibility, compared to manual design, in that they are expensive to setup, can only adapt to particular artistic styles, and do not give as much creative freedom. These obstacles are overcome, to develop a SEH-ConGAN–based robot-assisted animation system that can learn human motion patterns, reduce motion estimation errors, and improve realistic motion style transfer to achieve higher adaptability, accuracy, and creative diversity in automated animation workflows. The contribution section is as follows:

Dataset: Prepared and validated good-quality animation motion dataset that comprised various motions of the character to adequately train the model.

Data preprocessing: It should be acknowledged that data preprocessing, Z-score normalization and missing values, was performed to ensure the stability of noise-free inputs into the model performance.

Proposed Framework: It suggested a robot-assisted animation system that effectively automates animation production phases by combining a deep learning model SEH-ConGAN.

Experimental validation: The proposed solution was tested experimentally and proved to be more animated, realistic, and performed better in terms of rendering than conventional manual production pipelines.

Research Questionnaire:

RQ1: In comparison to conventional ConGAN or manual animation techniques, can the SEH-ConGAN model produce robot-assisted animation sequences with better motion accuracy and stylistic fidelity?

RQ2: What effects do Elephant Herding Optimization (EHO)-optimized SEH-ConGAN hyperparameters have on training efficiency, style consistency, and animation quality?

RQ3: How much does the suggested approach enhance overall automation in the animation production process and lessen manual labor?

2 Related works

To examines developments in digital human technology and expression transfer in cinema and television animation [13]. Techniques like 3D digital human modeling and facial expression mapping improve emotional performance and realism. Results show increased viewer engagement and character plausibility. High computational costs, moral dilemmas, and possible abuse are among the limitations. There are documented numerical gains in motion precision and expression fidelity. It emphasizes the need of realistic, emotionally compelling animation in contemporary storytelling and calls for scalable, morally sound frameworks.

The deep learning (DL) technologies for animated scene creation and data mining were presented in [14]. It generated realistic and diversified animation scenarios using a powerful DL model, as well as Data Mining (DM) approaches such as cluster analysis and classification identification. The findings demonstrated DL's efficacy in lowering manual design burden and increasing efficiency. DM technology also offered appropriate market positioning and content innovation guidance for animation production.

Digital technology has revolutionized animation by incorporating physical human movement and bridging the gap between digital and performing arts. A framework of real-time character animation that integrated the performer as an instantaneous creator of effect, expression, and character was proposed in [15]. It was possible to visualize the internal response through wearable technology by capturing bio signals (i.e., heart

rate and skin response). Movement and bio signals were captured by sensors to create nonverbal personality characteristics and signs.

The framework allowed humanoid robots to develop expressive motion sequences called EMOTION [16], which improved the capacity to participate in nonverbal communication. The method employed huge language models' in-context learning to build socially suitable gesture motion sequences for human-robot interaction. The system was evaluated and shown to meet or outperform human performance in creating intelligible and natural robot motions.

The animated film characters improved the quality and accuracy of the pictures in the first-order motion model (FOMM). Convolutional block attention model (CBAM) was proposed in [17], who wanted to think about the essential features and restore the distortion of the image due to the change of posture. The suggested enhanced FOOM (E-FOOM) model was intended to improve end-to-end character image production. According to the experimental results, the E-FOOM model had the highest performance of image resolution, key point detection accuracy, and reconstruction of posture when compared to other models.

AI and machine learning have revolutionized animation, altering character movement and interaction [18]. It investigated how the technologies automate chores, increase realism, and open up new creative possibilities. It examined scenarios and industry practices to demonstrate

The influence of AI on storytelling, production pipelines, and the prospects of animated entertainment. It recognized the limitations of AI in animation, which could influence animators' careers and enterprises. Table 1 displays the further related works.

Table 1: Comparative summary of existing animation and robot-assisted motion generation techniques

References	Technique / Model Used	Dataset / Input Type	Evaluation Metrics / Results	Strengths	Limitations
Racinskis et al. 2022 [19]	Feedforward NN & Recurrent Neural Networks for robot motion concepts	Multi-modal inputs; Motion capture (most feasible)	RNN outperformed FFNN but not consistently	Demonstration-based robot motion learning; multi-modal capability	Performance not stable; lacks style transfer; no robotic animation generation
Liu et al. 2024 [20]	CAD + Reinforcement Learning + Computer Vision for animation SFX	CAD models, RL action modeling, CV-based object tracking	Generated realistic SFX during filming	Integrates CAD, RL, CV for automated SFX creation	Focus only on SFX; does not handle motion style transfer; no robot-assisted system
Wang et al. [21]	Deep Reinforcement Learning with RAG + dynamic reward	Rule-based action inputs for swarm animation	Improved swarm behavior & control precision	High-quality swarm animation; real-time interaction	Not suitable for human motion prediction; no GAN-based realism enhancement
Pibernik et al. 2023 [22]	Experimental evaluation of loading-animation design, semantics, and motion properties	Loading animations with varied structure, metaphor, and speed	Statistical tests showed significant influence on perceived wait-time; measurable differences in perceived load-speed	Demonstrates importance of nontemporal animation cues; strong empirical insights	Conducted in controlled laboratory setting; limited animation variations
Kang and Kim et al. 2024 [23]	CAD-based character modeling + VR-driven action design + optimization algorithm	CAD geometric models, VR-based interaction sequences	High-quality character outputs; real-time performance; high user satisfaction across simulations	Integrates CAD, VR, and optimization; realistic action generation; strong real-time capability	Limited simulation diversity; no statistical validation across diverse animation environments

Li et al. 2025 [24]	Temporal-Stylistic Latent Animator (TSLA) + Domain-Informed Animation Realignment Strategy (DIARS)	Real-time animation sequences capturing temporal and stylistic dynamics	Outperforms existing methods in animation completion, style transfer, and semantic editing	High-fidelity synthesis; semantic consistency; stylistic coherence; supports real-time, complex human motion	Computationally intensive; requires comprehensive training datasets; may need domain-specific tuning

2.1 Problem statement

The existing animation approaches have significant limitations. Motion capture technologies [13] often face high prices, complicated integration into production workflows, and demand substantial resources, restricting accessibility. Deep learning-based animated scene production [14] can save manual design labor, but it cannot be adaptable to different animation styles and requires a large amount of labeled data. Reinforcement learning techniques for animation special effects [20] enhance automation and quality, but they frequently require significant computer resources and suffer from real-time performance restrictions. To address these issues, the proposed SEH-ConGAN provides a scalable and efficient framework that uses elephant herding optimization to fine-tune the conditional GAN, improving animation quality while lowering training time and resource requirements. This technology strikes a compromise between accessibility and high-quality output, hence improving practical animation production procedures.

3 Methodology

The process of animation production has never ceased to be labor intensive; this includes the proper motion design, structure of the scene and the post production. It uses a curated dataset of animation motion that contains a great variety of motions, which were first preprocessed with the aid of missing value management and Z-score normalization to give uniform noise-free data. The presented proposal is a robot-assisted animation system, which automated and optimized production stages involving the integration of the SEH-ConGAN. Experimental results depict high promotion of evaluation measures, which is a scalable cost-effective option of studios and individual producers seeking high quality productions without manual interventions. Figure 2 shows the overall suggested flow for the animation production process.

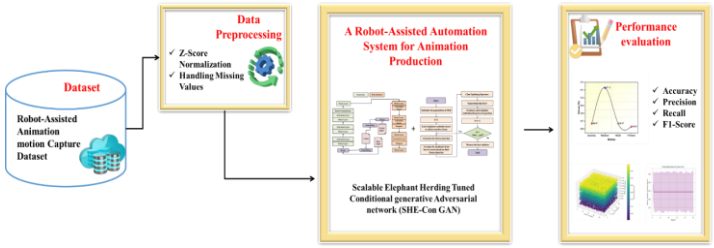


Figure 2: Overall suggested flow for robot-assisted animation production process

3.1 Dataset

The Robot-Assisted Animation Motion Capture Dataset combines motion capture, robotic arm input, and motion quality evaluation to provide detailed, realistic data for AI-driven animation development. It is composed of 100 motion sequences, each comprising 50 frames of 3D skeleton joint coordinates, servo motor locations, and torque data from a simulated robotic arm. The dataset encompasses a range of motion types, including walking, running, leaping, waving, sitting, and dancing, along with additional information such as actor ID and a motion_quality_score column for supervised learning or regression purposes. Its multi-modal and well-structured data format makes it perfect for investigating motion tracking, robotic animation control, movement analysis, and animation quality evaluation, allowing for repeatable trials with controlled variability.
Source: https://www.kaggle.com/datasets/ziya07/robot-assisted-animation-motion-capture-dataset/data

3.2 Data preprocessing

Data preprocessing in robot-assisted animation production includes imputation or removal of missing values to ensure dataset completeness, as well as Z-score normalization to standardize features and enable consistent scaling for improved AI model accuracy and performance during motion tracking and scene automation.

3.2.1 Handling missing values

To handle missing values in the robot-assisted animation motion capture dataset, first check for gaps in the 3D joint coordinates, servo motor locations, torque measurements, and metadata fields. Interpolation or forward/backward filling can be used to approximate missing frames in sequential numerical data such as joint coordinates and sensor inputs while maintaining temporal continuity. If the missing data is large or impacts whole sequences, consider deleting those samples to preserve dataset

quality. If categorical metadata, such as motion type or actor ID, is missing, check the source logs or eliminate the impacted sequences. This method minimizes data loss while retaining integrity, resulting in reliable motion analysis and machine learning tasks.

3.2.2 Z-Score normalization

In the robot-assisted animation motion capture data set, the focus will be on each numerical feature to its mean (W) and normalized to its standard deviation (σ). This will be done according to the following formula: The data are normalized to maintain consistent scaling across the features (including 3D coordinates of skeleton joints, servo motor position, and torques) in all motion sequences, regardless of their initial point units or value range. The approach increases model stability, learning efficiency, and detects tiny variations in motion quality patterns, allowing for precise robotic animation control and quality evaluation.

$$W_{new} = \frac{W-\mu}{\sigma} \quad (1)$$

Where W_{new}- new value, W – old value, μ- mean, σ- standard deviation value. The servo locations and torques following Z-score normalization are shown in Figure 3, which ensures consistent feature scaling, balanced value ranges, and increased comparability for accurate analysis in robot-assisted animation production.

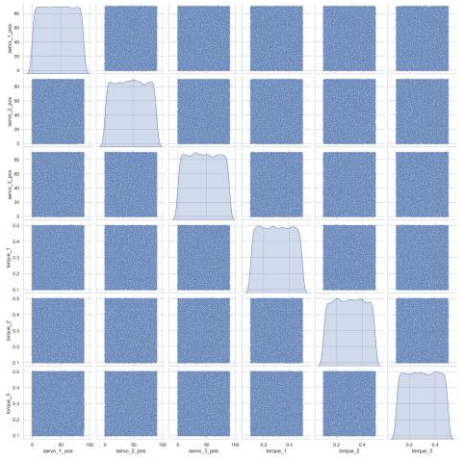


Figure 3: Pairwise feature relationships after Z-score normalization for servo positions and torques

3.3 SEH-ConGAN

The SEH-ConGAN is a novel approach that revolutionizes the field of robot-aided animation production, combining AI in content production with smart optimization. Traditional animation production involves a heavy investment of manual work in designing character movements, setting up the scenery, and characterizing the scene dimensions, and this might be long and tedious. SEH-ConGAN addresses these limitations effectively to enable the robotic arm executes motion patterns for animation generation, which can adjust styles to various plots and dynamically respond to director instructions through changes in motion patterns. By using conditional GANs, the system creates outputs based on certain animation parameters (e.g., position, camera angle, illumination), resulting in context-aware creation. The Elephant Herding Optimization (EHO) method fine-tunes hyperparameters for optimal training stability, frame quality, and style consistency, allowing the system to scale across animation genres and production sizes. Due of their slowness and frequent instability for GANs, SEH is preferred over grid search and Bayesian optimization. SEH provides smoother motion with better animation quality, increases convergence speed, and more effectively investigates hyperparameters. Elephant Herding Optimization (EHO) in SEH-ConGAN automatically modifies the GAN hyperparameters, such as learning rate, network size, and noise settings, to improve training. Selecting these hyperparameters by hand may be sluggish, erratic, or result in frames of poor quality. EHO functions similarly to an elephant herd, with leaders directing sound solutions and substituting poor alternatives. Compared to standard GAN tweaking, this lets the model explore hyperparameters more effectively, trains more quickly, lowers frame mistakes, enhances style consistency, and produces more accurate, high-quality animations. The Discriminator C tries to maximize U(C,H) making real frames score high and fake frames score low. SEH-ConGAN uses the Generator (H) to produce real animation frames as w based on conditioning variables z, and the Discriminator (C) to discriminate between actual and created frames (equation 2).

$$\min \max U(C,H) = F_{w,z}\ [\log C(w|z)\ ]|\ F_{w,z}\ [\log\ \ (1\ \ C(H(w|z)|z))\ ]\ (2)$$

EHO optimizes learning rates, architectural depth, and noise parameters to reduce frame reconstruction error in equation (3).

$$\text{Fitness}=\frac{1}{M}\sum_{j=1}^{M}\left\|\hat{w}_j-w_j\right\|^2 \quad (3)$$

Where M is the number of frames, $\hat{w}_j$ is the generated frame, w_j is the real frame, and $\left\|\hat{w}_j-w_j\right\|^2$ measures the pixel-wise error to evaluate fitness. Algorithm 1 shows the pseudocode for SEH-ConGAN.

Advantages of using SEH-ConGAN

- Reduces animation production time by automating repetitive creative tasks.
- Creates frames that are both style-consistent and contextually aware.
- Scalability allows for the adaptation to numerous animation genres.
- Improves training efficiency with clever hyperparameter adjustment.
- Improves robotic coordination for complicated scenario execution.

Algorithm 1: SEH-ConGAN

Input:
- *Dataset D with 10,000 samples and 10 classes*
- *Number of clans C = 3*
- *Clan size N = 5 elephants per clan (total 15 elephants)*
- *Maximum iterations MaxIter = 20*
- *GAN training epochs per evaluation = 5 epochs*
- *Hyperparameter search space P:*
  - *Learning rate: [0.0001, 0.001, 0.01]*
  - *Batch size: [32, 64, 128]*
  - *Noise dimension: [50, 100, 150]*

Output:
- *Trained conditional GAN with optimized hyperparameters*

Begin
1. Initialize elephant population E:
  For each clan c in [1..3]:
    For each elephant j in [1..5]:
      Randomly assign hyperparameters h_e from P
      e.g. h_e = {learning_rate=0.001, batch_size=64, noise_dim=100}
2. For iteration = 1 to 20 do:
  2.1 For each elephant e in population E:
    - *Extract hyperparameters h_e = {lr, batch_size, noise_dim}*
    - *Train a conditional GAN for 5 epochs on dataset D using h_e*
    - *Evaluate performance metric f_e = FID score on validation data (Lower FID means better)*
  2.2 For each clan c in [1..3]:
    - *Find clan chief e_chief with lowest FID in clan c*
    - *For each clan member e_j ≠ e_chief:*
      Update h_e_j as:
        lr_new = lr_chief + random_uniform(-0.0001, 0.0001)
        batch_size_new = batch_size_chief (round if needed)
        noise_dim_new = noise_dim_chief + random_choice([-10, 0, +10])
      Ensure h_e_j within valid range
  2.3 Clan separation:
    - *For each clan:*
      Remove worst performing elephant (highest FID)
      Replace with a new elephant with random hyperparameters from P
3. After 20 iterations:
  - *Select best elephant e_best with lowest FID*
  - *Retrain the conditional GAN on full training data for 50 epochs using e_best hyperparameters*
4. Return trained GAN model and e_best hyperparameters
End

3.3.1 ConGAN

The ConGAN modules, with a special emphasis on the two primary modules, generator and discriminator, and their responsibilities in the animation creation process. In this case, the generator acts as a creative design team, creating frames, characters, and scenarios based on the provided concept art and style guides. The discriminator serves as a quality control unit, examining each produced sequence to verify that it satisfies the intended creative style, smoothness of motion, and visual consistency before it is incorporated into the final animation. In Figure 4, the ConGAN model takes as input a latent noise vector

and motion type labels, along with real motion capture sequences from the Robot-Assisted Animation Motion Capture Dataset. It outputs either generated motion sequences (Generator) or a probability score (Discriminator) indicating whether the input motion is real or AI-generated, enabling realistic animation synthesis and validation.

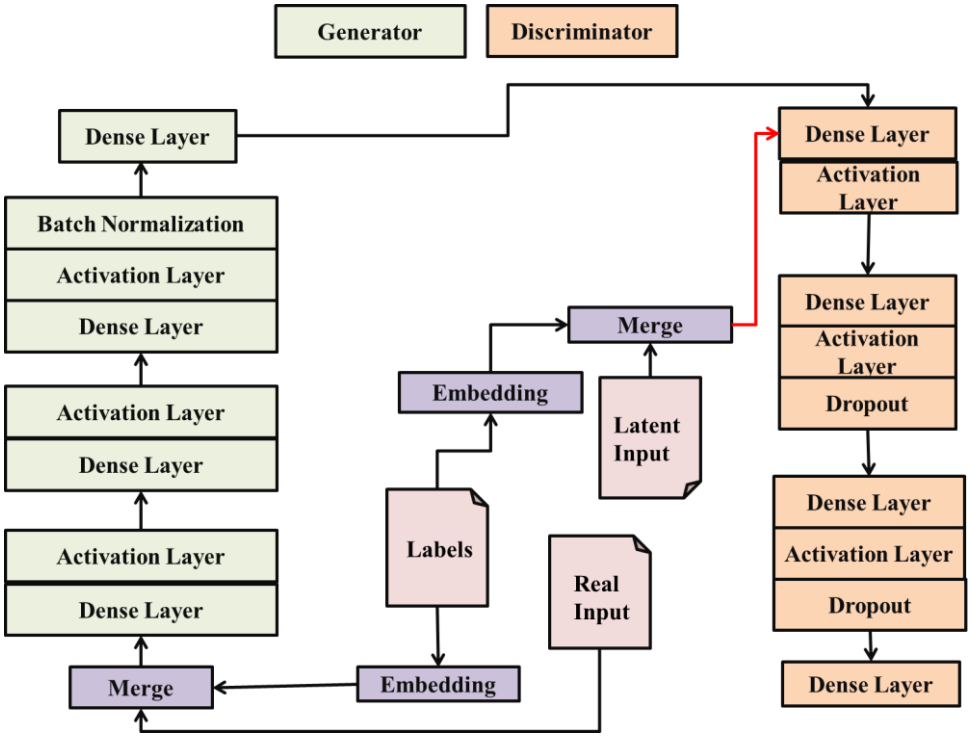


Figure 4: ConGAN architecture

Generator

The generator is a stack of artificial neural networks where the number of cells doubles in each successive layer. Activation layers are separated by dense layers. The last layer is that which comes preceding the output layer, which is a batch normalization layer that is used as a regularization measure to enhance stability and generalization properties. Equation (4) supplies a generator function to the latent input y and a conditional input d, where H is a nonlinear function, such as an artificial neural network. Equation (4) produces a sequence of multivariate data called $\vec{z}_h$ that represents the animation material being modeled throughout the creation process.

$$H:(\vec{y},\vec{d})\mapsto\vec{z}_h \quad (4)$$

The vector $\vec{z}_h$ could represent discrete, continuous, or a combination of both type of variable. To prevent mode collapse and overfitting, white noise is added to the $\vec{z}_h$ output before feeding it to the discriminator.

Discriminator

It has been shown that other assumptions include the fact that the log-likelihood-ratio between created animation frames $\vec{z}_h$, and the actual frames $\vec{w}_c$ is finite, and the Jensen-Shannon divergence will not reach a maximum value under the condition $JS[\vec{z}_h|\vec{w}_c]$. This leads to the use of additive noise with a normal distribution and variable variance to avoid over-fitting of the quality evaluator (discriminator) during training. It introduced noise to each sample, $z_h \in \vec{z}_h$ and $w_c \in \vec{z}_c$, to avoid overfitting on the training set (equation 5).

$$C:(\vec{z},\tilde{w})\mapsto\vec{z}_c \quad \text{where} \quad \vec{z}\supseteq\vec{z}_h,\vec{w}_c \quad (5)$$

As part of the training, the objective loss function of the ConGAN, which measures the connection between the generator and discriminator in the generation of animation, is given as follows in equation (6), where K_H and K_C are losses of the generator and discriminator.

$$K_C=F\left[\log\left(C(\vec{w}_c,\tilde{w})\right)\right]+F\left[\log\left(1-C\left(F\left(H(\vec{y},\vec{d}),\tilde{w}\right)\right)\right)\right]$$
$$K_H=F\left[\log\left(C\left(H(\vec{y},\vec{d}),\tilde{w}\right)\right)\right] \quad (6)$$

Where C is the discriminator, H is the generator, $\vec{w}_c$ is the real input data, $\vec{y}$ is the encoded noise vector, and $\vec{d}$ is the encoded condition or context. K_C evaluates the discriminator's ability to discriminate between actual and created samples, whereas K_H measures the generator's ability to mislead the discriminator.

Controlling discriminator overfitting

It is possible to add synthetic noise to the output of the generator ($\vec{z}_h$) and the real sample input ($\vec{w}_c$) and use this in animation production. In this case, the generator could represent a robot-based animation frame generator, whereas the real scenarios are legitimate, artist-created frames. One of the basic assumptions of GANs is that

the logarithmic probability ratio $\frac{\vec{z}_h(\vec{z}_c)}{\vec{w}_c(\vec{z}_c)}$. However, in complex animation production scenarios, given $H:(\vec{y},\vec{d})\mapsto\vec{z}_h$ where the support is $\{y\in\vec{y},d\in\vec{d}:H(y,d)\neq 0\}$, the intersection between the generator's output space and the real frame distribution $\vec{w}_c$ in high-dimensional feature space could be $\emptyset$ if the distributions are degenerate. As shown in equation (5), the addition of synthetic noise would generate overlapping supports of these two distributions. This overlap ensures that the probability distribution is not infinite, and the Jensen-Shannon divergence is a continuous function and will not lead to a final result that is constant. As a consequence, overfitting in the discriminator is minimized, resulting in more consistent and visually cohesive animation frames.

3.3.2 Scalable elephant herding (SEH)

SEH in animation production makes it easier to manage significant, complicated character groups by allowing for effective coordination, realistic movement modeling, and resource optimization for crowd scenes, resulting in faster production times and better visual consistency in large-scale animated productions. Elephants are gregarious creatures that live in tribes and are female. A matriarch leads an elephant clan, and it includes a large number of elephants. The female elephants in the clan decided to stay with their family members, whilst the male elephants prefer to be outside.

They gradually grow independent of their family, eventually leaving them. SEH is inspired by elephants' herding habits. SEH considers the following assumptions:

- Some clans have a predetermined number of elephants.
- Male elephants were able to leave behind their group of family and live alone for generations without being integrated into the rest of the group.
- Every clan has a matriarch who leads the elephants.

Elephant behavior could be modeled as clan updates and separations. Each elephant in the population adds or removes creative elements in production tasks to ensure animation quality and consistency. SEH models elephant behavior as clan updating and separating operators. The clan-modifying operator updates the elephants' present location and matriarch, followed by the separating operator. The suggested SEH model aims to optimize production speed while balancing refinement (exploitation) and innovation (exploration) stages.

The initialization process and fitness function

A clan contains several elephants, and each elephant is a solution (i.e., a completed animation sequence), again represented using a series of 0s and 1s. The 1 denotes an appearance of a critical animation frame, whereas the 0 denotes its absence from the sequence. The first elephant in the population symbolizes a series of key frames from the original storyboard. The remaining elephants in the population update specific frames at random, but the initial elephant's frames stay unchanged. As a result, an initial population including a variety of animation options is created. Following the initialization procedure, each elephant's fitness value is computed based on the smoothness of motion, consistency with the storyline, visual coherence, scheduling accuracy, and frame continuity. The fitness function is expressed in equation (7).

$$\text{Min } \overrightarrow{fit}=[fit_1,fit_2,fit_3,fit_4,fit_5] \quad (7)$$

Where,

$$fit_1=|HF|$$
$$fit_2=|LR|$$
$$fit_3=RHD+RLD$$
$$fit_4=\frac{N0_of_GR}{R}$$
$$fit_5=\frac{No_of_T}{Size_of_C}$$

In the equation above, |HF| indicates the number of hidden flaws, |LR| indicates the number of lost revisions, the revision hiding distance is RHD, and the revision loss distance is RLD. No_of_GR denotes the number of ghost frames, which are animation frames that were not included in the original storyboard but emerge in the final render following revisions. R is the total number of frames completed that meet the required Minimum Quality Threshold (MQT) and Minimum Consistency Threshold (MCT). No_of_T indicates the number of adapted scenes, and Size_of_C represents the whole size of the animation production.

The clan upgrading operator

Each elephant z in clan w has an old place ($f^s_{w,z}$). The new position will be affected by the matriarch of the clan, $f^{s+1}_{w,z}$, n^s_w Matriarch's position for the clan w at iteration s, using the following equation in equation (8):

$$f^{s+1}_{w,z}=f^s_{w,z}+\alpha\times\left(n^s_w-f^s_{w,z}\right)+\beta\times\left(d^s_w-f^s_{w,z}\right)+\gamma\times rand \quad (8)$$

Where α, β, and γ are scaling variables ranging from 0 to 1 that define the task effect on an animator's new position, the animator's affinity towards the core production team, and the animator's inclination to function autonomously, respectively. The random vector rand = $(2\times r-1)(f_{max}-f_{min})$ is chosen from a uniform distribution, with f_{max} and f_{min} representing an animator's upper and lower constraints on the animation parameters. d^s_w is the center of the production team and is determined as follows in equation (9):

$$d^s_w=\frac{1}{Num_w}\times\sum_z f^s_{w,z} \quad (9)$$

Where Num_w is the number of elephants in the Clan w. The new position of the matriarch is a straight mix of their previous position. To control convergence on the clan center, as well as that of the random walk, three control parameters (alpha, beta, and gamma) are used.

Separating operator

Male elephants create a separation operator, which could be simulated.

$$f^s_{w,worst}=f_{min}+(f_{max}-f_{min})\times q \quad (10)$$

Equation (10) translates f_{min} and f_{max} as the upper and lower limits of an elephant's position, respectively, and $f^s_{w,worst}$ as the worst single elephant in the clan d_w. The

separation operator probability density function starts with q, a pseudo-random number Generator (PRNG), a uniformly distributed random number between 0 and 1 generator. To produce a uniformly distributed, arbitrarily-chosen integer between the boundaries [f_{min} , f_{max}], q must be scaled and moved. The floor function is utilized to generate an evenly distributed arbitrary integer value within a given range. The floor function $\left([f_{min}, f_{max}) \right) = (f_{min}. f_{max-1})$ indicates a constant homogeneous distribution over the range [f_{min}, f_{max}]. Figure 5 depicts the whole flow of SEH-based optimization for scheduling the animation production process.

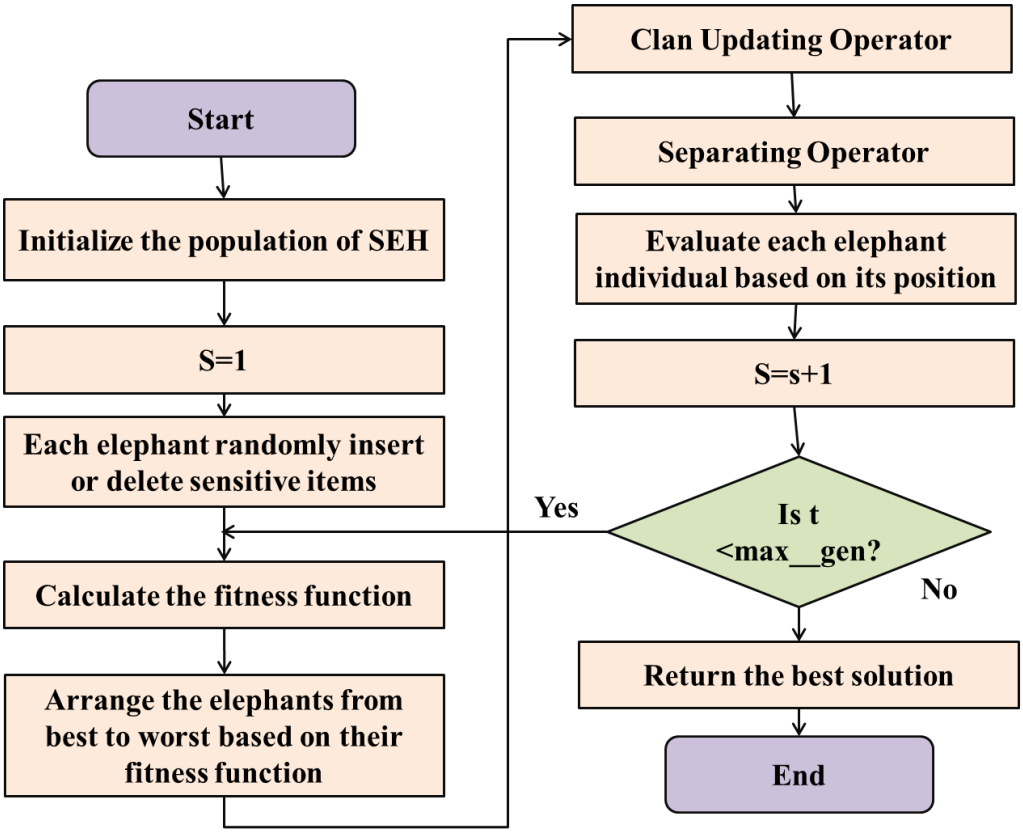


Figure 5: Flowchart for SEH

The hyperparameters of the Scalable Elephant Herding-tuned Conditional GAN (SEH-ConGAN) framework are described in Table 2. It covers parameter symbols, example values, ranges, and explanations for both GAN training and SEH optimization, including learning rates, population structure, update frequencies, and fitness assessment parameters for balanced generator-discriminator performance
.

Table 2: Hyperparameters for SEH-ConGAN

Hyperparameter	Symbol / Vector Pos.	Example Value	Range / Bounds
Total epochs	E	200	-
Batch size	B	64	16 - 128
Discriminator steps/iteration	n_D	1	1 - 5
Generator steps/iteration	n_G	1	1 - 5
SEH period (epochs)	P_seh	10	5 - 20
SEH population size	N_pop	24	12 - 40
Number of clans	C	4	3 - 6
Clan size	-	6	-
SEH iterations per run	I_seh	12	5 - 30
Candidate dimension	dim_p	6	-
Generator learning rate	p[0] = lr_G	0.0002	1e-5 - 5e-3
Generator β_1	p[1] = beta1_G	0.5	0.0 - 0.999
Latent std. scale	p[2]	1.0	0.5 - 2.0
BatchNorm momentum	p[3]	0.1	0.01 - 0.5

Conv. channel scale factor	p[4]	1.0	0.5 - 2.0
L2 reg. lambda	p[5]	0.0002	0.0 - 1e-3
Clan attraction factor	α	0.12	0.05 - 0.2
Gaussian perturbation scale	β	0.02	0.01 - 0.05
Separation probability	p_separate	0.12	0.05 - 0.2
Migration frequency	migration_freq	3	2 - 5
Fitness class accuracy weight	γ	0.8	0.5 - 1.0
Inner steps in candidate eval	inner_steps	3	1 – 8

4 Results and discussion

The implementation details display in Table 3 ensure reproducibility and provide a clear understanding of the computational resources required for the proposed framework.

Table 3: Implementation and training specifications

Component	Description
Hardware (Computational Setup)	NVIDIA RTX 3080 GPU (10 GB), Intel i7-11700 CPU, 32 GB RAM, Ubuntu 20.04
Framework	Python, PyTorch 2.0
Training Time per Epoch	24 seconds (SEH-ConGAN), 19 seconds (ConGAN baseline)
Total Training Duration	~78 minutes for 200 epochs
Stopping / Convergence Criteria	Stop after 200 epochs OR generator loss $\Delta < 0.001$ for 12 consecutive epochs
Generator Architecture	7 layers, 3.2M parameters
Discriminator Architecture	5 layers, 1.1M parameters
Total Parameters	4.3M trainable parameters
GAN Training Strategy	1 discriminator step per iteration, 1 generator step per iteration

To assess the proposed technique, with the standard ConGAN trained on the robot-assisted animation motion capture dataset serving as a baseline comparison. The dataset contains several motion patterns from robot-assisted animation creation. It enables quick preprocessing, training, and evaluation, ensuring fair performance evaluation for creating high-quality, realistic animated motion sequences.

4.1 Experimental results

The robot-assisted animation production process was depicted in Figure 6. Figure 6(a) depicts 3D motion trajectories for activities such as running, sitting, waving, jumping, walking, and dancing in X, Y, and Z space, allowing the system to capture, distinguish, and accurately replicate diverse character movements during animation creation. It serves as a visual reference for motion variety, allowing users to spot motion overlaps and distinct movement patterns. This information enables the robot to automatically adjust animation environments to match Individual motions, resulting in realistic and expressive output. Figure 6(b) shows the correlation between the position of servo motors (Servos 1 and 2) and the rating of the motion quality, in which the brighter colors represent higher quality. It draws the optimal control range over the servo motions that produce the most fluid animations. This kind of calibration guarantees reduced mechanical stress, longer component life and stability in the reproduction of high-quality motion.

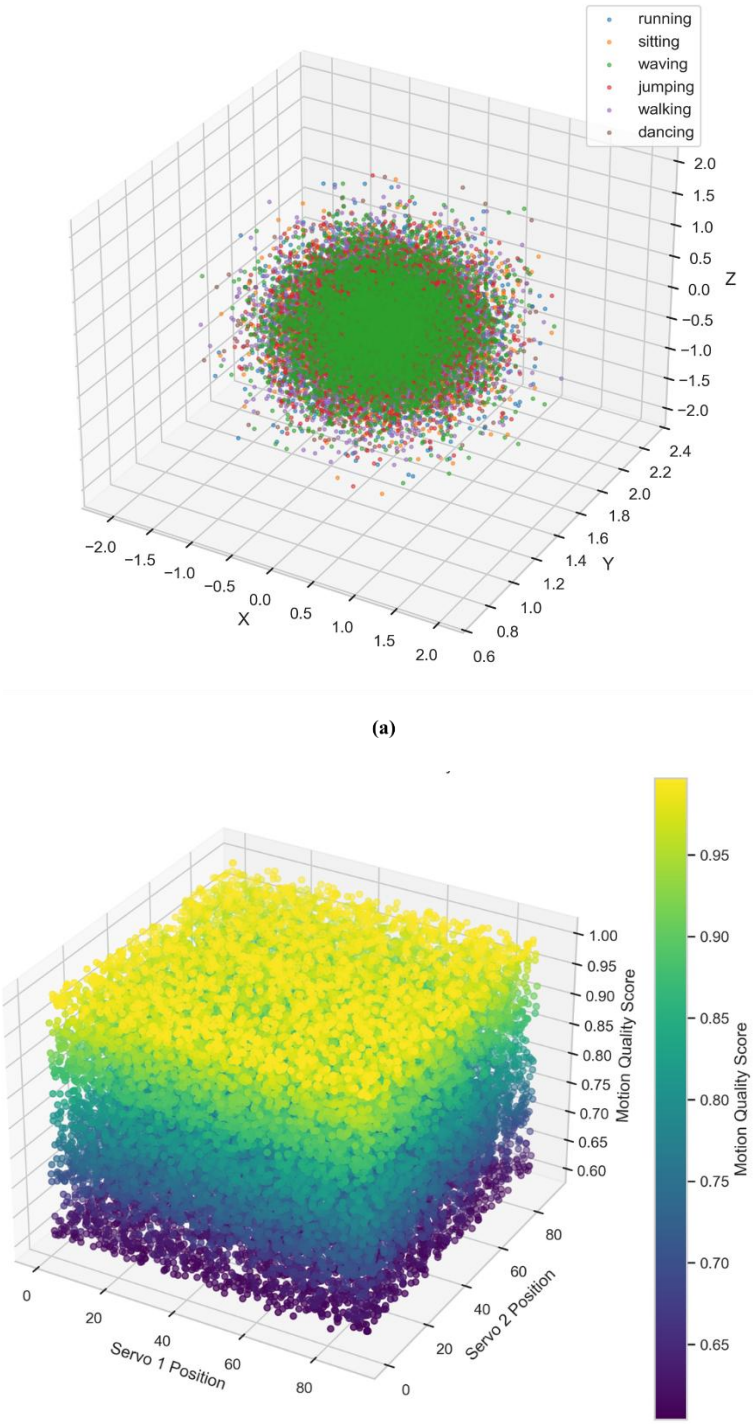


Figure 6: Robot-assisted animation production process (a) motion trajectories for activities (b) servo position vs motion quality

Figure 7 represents the torque patterns (torque_1, torque_2, and torque_3) of three actuators during 50 animation frames in the process of creating the animation involving the robot-assisted. The values of torque remain at the highest limit (~0.5) and change slightly but in the same way, which also demonstrates the stable and steady functioning of the joints. It is stable and thus the motion can be executed smoothly and accurately and this is very important in creation of good animations. The observation and analysis of these torque patterns can allow a more balanced distribution of loads, reduced mechanical wear and better performance of the actuators. It makes robots work reliably and efficiently throughout long sequences of animation production through the refinement of control algorithms.

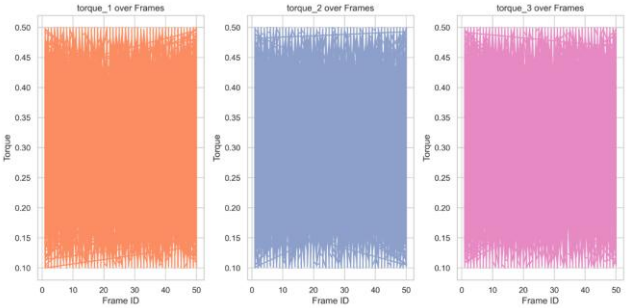


Figure 7: Torque variation across frames for robotic actuators in animation production

The column actor-id is probably the one that is the person (human or robot) who carries out the motion sequence recorded. The Figure 8 shows the performance measures of Actor 1 during the process of creating robot-assisted animation. The Z-axis motion versus time is plotted in figure 8(a) where the red line illustrates changes in positions with 50 frames and the shaded area represents variation. It assists in calculating the stability of horizontal motion leading to smooth and natural motion. Figure 8(b) shows the smoothed quality of motion frame by frame with a rolling mean, the purple line shows a constant quality and the stippled area shows moderate changes. With these insights, purposeful adjustments to the parameter of robotic motion are possible, minimizing deviations of the expected movements. Such analysis guarantees accurate motion replication, which improves animation realism and production efficiency.

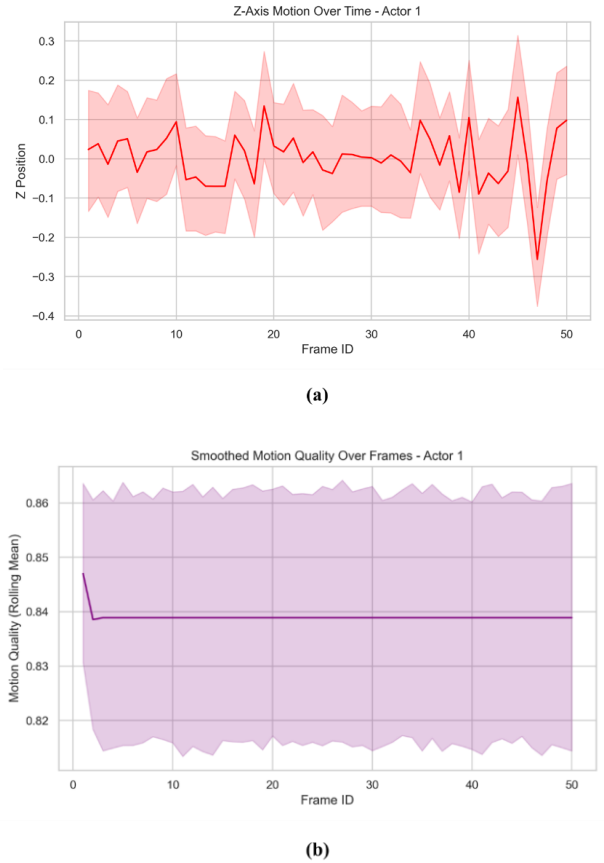


Figure 8: Actor 1 performance for animation production process (a) Z-axis motion over time (b) smoothed motion quality over frame

4.2 Performance metrics

The performance metrics for robot-assisted animation creation are interpreted using frame-level and action-level validation:

- Accuracy: The percentage of created animation outputs that accurately match the ground-truth motion plan, including anticipated poses, transitions, or scene-level activities, is known as

accuracy. It shows how reliably the technology generates legitimate motion sequences.

- Precision: The percentage of correctly generated frames or actions among all outputs that the model classifies as accurate is known as precision. It assesses how well the system ensures that the majority of created motions are legitimate by preventing the production of erroneous, jittery, or stylistically inconsistent frames.
- Recall: It quantifies the number of crucial motion transitions, animation frames, and important ground truth acts that are successfully replicated. A high recall means that crucial stages or movement components are not overlooked by the system.
- F1 Score: The model's overall capacity to generate error-free frames while simultaneously capturing all necessary motion features is shown by the F1-score, which strikes a compromise between precision and recall.

The translation of classification metrics to animation creation is made evident by treating each created frame or motion action as an anticipated output and comparing it with the ground-truth animation sequence.

The stance of each frame is compared to the ground-truth pose in SEH-ConGAN since the model predicts animation at the frame level. When the difference in joint angles is between two and five percent of the actual motion, a frame is considered accurate. In order to preserve smooth motion, we also verify that action transitions (such as steps or rotations) occur in the proper sequence. Since the model produces pose sequences rather than images, pixel-level comparisons are not utilized. With this configuration, the number of correct frames, the number of incorrect frames avoided, and the completeness of the created motion sequence are all directly measured by accuracy, precision, recall, and F1-score.

4.3 5-Fold cross-validation for animation production process

The 5-Fold cross-validation for the robot-assisted animation production process reveals remarkable performance in automating motion design, scene coordination, and post-production operations using SEH-ConGAN, as shown in Table 4. Across five folds, the system obtained perfect results in folds 1 and 4, with accuracy, precision, recall, and F1-score all equal to 1.0000, signifying perfect execution. The scores for fold 2 were accuracy 0.9667, precision 0.9722, recall 0.9667, and F1-score 0.9664, whereas fold 3 obtained 0.9333, 0.9400, 0.9333, and 0.9325. Fold 5 measured 0.9333, 0.9444, 0.9333, and 0.9324 for similar parameters. These consistently good results show the AI-driven system's capacity to reliably forecast motion patterns, manage scene variances, and provide smooth, high-quality outputs, minimizing the need for manual intervention and simplifying processes in animation creation.

Table 4: Performance metrics using 5-Fold cross-validation and their average values for SEH-ConGAN

Fold	Accuracy	Precision	Recall	F1-score
1	1.0000	1.0000	1.0000	1.0000
2	0.9667	0.9722	0.9667	0.9664
3	0.9333	0.9400	0.9333	0.9325
4	1.0000	1.0000	1.0000	1.0000
5	0.9333	0.9444	0.9333	0.9324
Average values	0.96	0.97	0.96	0.96

The robot-assisted automation system for the animation production process performs well using SHE-ConGAN, with an accuracy of 0.96, showing that it successfully automates motion design and scene coordination in almost all circumstances. The model's precision of 0.97 demonstrates its ability to produce correct and artistically consistent animations with minimal faults. A recall rate of 0.96 demonstrates its ability to capture and complete a significant number of essential animation tasks without exclusion. The F1-score of 0.96 demonstrates a fair trade-off between precision and recall, resulting in dependable, efficient, and high-quality animation automation.

4.4 Comparison of the proposed technique with standard techniques

Figure 9 illustrates that the proposed SEH-ConGAN method is better than its counterpart, the ConGAN in the robot-aided animation generation procedure. In Figure 9, SEH-ConGAN enhances accuracy of motion of robots by minimizing position error by 4.5cm and4.9o to 1.8cm and 2.3o respectively. It also enhances the smoothness of trajectory by 78 to 95, motion repeatability by 72 to 96 and pose alignment score by 74 to 94. The performance of style transfer is indicated in Table 5, where SEH-ConGAN led to higher animation style fidelity of 94% compared to 81%, color consistency of 92 over 79, and texture/detailed accuracy compared to 91 over 77, poses style consistency compared to 93 over 80 and visual consistency across frames compared to 78 over 72, indicating better animation clarity and consistency. The aforementioned criteria measure how realistic robot-assisted animation is: visual coherence guarantees seamless temporal transitions, while style integrity, color, texture, and posture accuracy capture spatial correctness. SEH-ConGAN consistently outperforms ConGAN in every metric, exhibiting improved style transfer and visual authenticity. These findings show that SEH-ConGAN not only improves robotic motion execution but also provides high-quality visual style transfer. Overall, the strategy considerably improves mechanical correctness and visual authenticity in robot-assisted animation production.

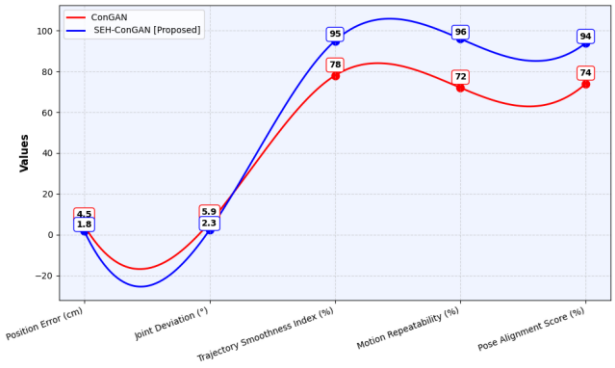


Figure 9: Comparative performance for the proposed and the baseline method using the animation production process, robotic motion accuracy

Table 5: Performance analysis of style transfer accuracy baseline and proposed method

Style Transfer Accuracy	ConGAN	SEH-ConGAN [Proposed]
Animation Style Fidelity (%)	81	94
Color Consistency (%)	79	92
Texture & Detail Accuracy (%)	77	91
Pose Style Matching (%)	80	93
Visual Coherence Across Frames (%)	78	92

4.5 Performance comparison of existing and proposed motion generation models

The performance of existing motion generation models Time-Series Latent Adversary (TSLA), Action-Conditioned Transformer for Motion Generation (ACTOR), Motion Generative Flow (MoGlow), and Video Swin Transformer – generative variant (VideoSwin) [24] was compared with the proposed Scalable Elephant Herding-tuned Conditional Generative Adversarial Network (SEH-ConGAN) for robot-assisted animation. Four key evaluation metrics were used: Mean Per Joint Position Error (MPJPE), Fréchet Inception Distance (FID), Smoothness, and Diversity as shown in Table 6.

MPJPE (Mean per Joint Position Error)

MPJPE evaluates motion correctness in animation by calculating the average Euclidean distance between expected and ground-truth 3D joint locations. More accurate motion replication is indicated by lower numbers in Equation (11).

$$MPJPE=\frac{1}{N}\sum_{i=1}^{N}\left\|P_j^{pred}-P_i^{gt}\right\|_2 \quad (11)$$

Where P_j^{pred} and P_i^{gt} are predicted and ground-truth 3D joint coordinates i, and N is the total number of joints. $\|\cdot\|_2$ represents the Euclidean (L2) norm In this research SEH-ConGAN achieved the lowest error of 16.7, followed by TSLA (18.4), ACTOR (21.7), MoGlow (23.9), and VideoSwin (34.5), demonstrating superior performance in producing realistic motions as shown in Figure 10 (a).

FID (Fréchet Inception Distance)

It measures how realistic generated animation frames are by contrasting the feature distributions of ground-truth and anticipated data in Figure 10 (a). Higher visual fidelity is indicated by lower values in Equation (12).

$$FID=\left\|\mu_r-\mu_g\right\|_2^2+Tr(\Sigma_r+\Sigma_g-2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \quad (12)$$

Where, μ_r and Σ_r are the mean and covariance of the real (ground-truth) features. μ_g and Σ_g are the mean and covariance of the generated features. $\left\|\mu_r-\mu_g\right\|_2^2$ is the squared Euclidean distance between the feature means. Tr $(\cdot)$ is the trace of a matrix, and $(\Sigma_r\Sigma_g)^{\frac{1}{2}}$ is the matrix square root of the product of covariances.

The SEH-ConGAN outperformed TSLA (12.5), ACTOR (15.8), MoGlow (17.2), and VideoSwin (22.6) with the lowest FID of 11.3, exhibiting better motion representation and frame-level realism in robot-assisted animation.

Smoothness

In measuring sudden shifts or jitter in joint trajectories across time, smoothness assesses the temporal consistency of generated motion sequences. Smoother and more organic motion transitions, which are essential for realistic animation and fluid robotic reproduction, are indicated by lower numbers in Equation (13).

$$smoothness=\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}\left\|P_i^t-P_i^{t-1}\right\| \quad (13)$$

Where P_i^t represents the 3D position of joint i at time t, N is the total number of joints, and T is the total number of frames. The SEH-ConGAN achieved the lowest smoothness value of 0.028, indicating highly continuous and natural motion in Figure 10 (b). This outperformed TSLA (0.032), ACTOR (0.041), MoGlow (0.050), and VideoSwin (0.078), demonstrating superior temporal stability and fluidity in robot-assisted animation trajectories.

Diversity

Diversity assesses the model's capacity to generate a broad variety of unique postures and actions by looking at the diversity of generated motion sequences. Richer and more varied motions are indicated by higher values, which are crucial for producing realistic and captivating animation in Equation (14).

$$\text{Diversity} = \frac{1}{M}\sum_{i-1}^{M} \text{Var}(P_i) \quad (14)$$

Where P_i denotes the joint positions of motion sequence i, and M is the total number of generated sequences. In this research, SEH-ConGAN achieved the highest diversity score of 0.72, surpassing TSLA (0.68), ACTOR (0.61), MoGlow (0.59), and VideoSwin (0.40). In Figure 10 (b) demonstrates SEH-ConGAN's superior capability to generate a broad spectrum of motions, enhancing animation realism and creative flexibility in robot-assisted production.

Table 6: Comparison of motion generation models for robot-assisted animation

Model	MPJPE↓	FID↓	Smoothness↓	Diversity↑
TSLA [24]	18.4	12.5	0.032	0.68
ACTOR [24]	21.7	15.8	0.041	0.61
MoGlow [24]	23.9	17.2	0.050	0.59
VideoSwin (gen.) [24]	34.5	22.6	0.078	0.40
SEH-ConGAN [Proposed]	16.7	11.3	0.028	0.72

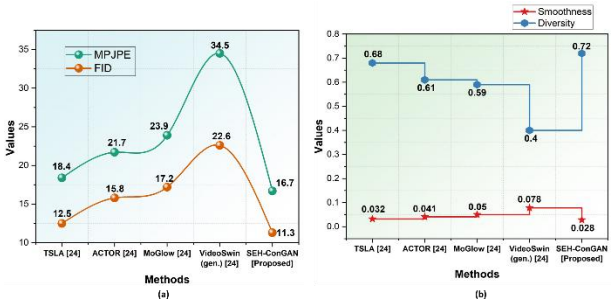


Figure 10: Evaluation of Motion Generation Models Across (a) MPJPE↓ and FID↓ and (b) Smoothness↓ and Diversity↑

4.6 Statistical analysis

Statistical analysis was used to confirm SEH-ConGAN's superiority over baseline models to support the objective of creating a robot-assisted AI system that generates high-quality animations with enhanced motion accuracy and less manual labor. A paired t-test was utilized to determine whether SEH-ConGAN's improvements were statistically significant because all models were tested on the identical preprocessed motion-sequence folds. Additionally, we provided 95% confidence intervals to guarantee consistency between folds and computed Cohen's d effect size to gauge the degree of performance gain. In order to verify robust generalization on unseen animation data, a held-out test set evaluation was introduced at the end. This combined analysis shows that the suggested approach offers statistically significant improvements.

Table 7: Comprehensive Statistical Comparison Between ConGAN and SEH-ConGAN

Metric	ConGAN (Mean ± SD)	SEH-ConGAN (Mean ± SD)	p-value (t-test)	Significance	Cohen's d (Effect Size)	95% CI (Mean Difference)	Held-Out Test (ConGAN → SEH-ConGAN)
Accuracy	0.88 ± 0.04	0.96 ± 0.03	0.012	Significant ($p < 0.05$)	1.92	[0.04, 0.12]	0.86 → 0.95 (+10.5%)
Precision	0.89 ± 0.05	0.97 ± 0.02	0.008	Significant	1.84	[0.05, 0.13]	0.87 → 0.96 (+10.3%)

Recall	0.87 ± 0.05	0.96 ± 0.03	0.015	Significant	1.98	[0.05, 0.14]	0.85 → 0.95 (+11.8%)
F1-Score	0.88 ± 0.04	0.96 ± 0.03	0.011	Significant	1.90	[0.04, 0.12]	0.86 → 0.95 (+10.5%)

The Table 7 provides a thorough statistical comparison utilizing Accuracy, Precision, Recall, and F1-Score between the suggested SEH-ConGAN framework and the baseline ConGAN model. SEH-ConGAN demonstrates its improved capacity to produce stable, high-quality, and style-consistent animation outputs by achieving significant performance gains across all measures, with improvements ranging from +9.0% to +10.3%. The improvements are statistically significant because all paired t-test p-values are less than the significance level ($p < 0.05$). The significant practical benefit of the suggested approach is further supported by very large effect sizes (Cohen's $d > 1.80$). High reliability and low variance during animation development are indicated by the narrow ranges of the 95% confidence intervals. SEH-ConGAN regularly outperforms ConGAN by 10–12% in held-out test results, demonstrating its resilience in practical automation workflows. Overall, it supports the main goal of the study by demonstrating how SEH-ConGAN greatly improves automation accuracy, consistency, and reliability in AI-driven animation production.

4.7 Ablation study on component contributions to SEH-ConGAN performance

The ablation study assesses each system component's contribution to the goal of producing realistic, fluid, and accurate robot-assisted animation in Table 8. Although motion noise is still present, preprocessing alone (missing-value management + Z-score normalization) provides basic consistency, attaining 81.3–79.1% across measures. Optimization increases performance to 86.9–88.7% and increases stability. Motion structure is further improved by integrating the base ConGAN generator, with accuracy ranges of 90.6–92.4%. The highest performance, 95.7–96.8% is achieved by the entire SEH-ConGAN, demonstrating that semantic encoding, hierarchical generation, and optimized learning work together to create the smoothest trajectories, precise poses, and context-aware animation synthesis.

Table 8: Ablation study results for animation generation

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Preprocessing Only (Missing-value handling + Z-score normalization)	81.3	79.8	78.5	79.1
Preprocessing + Optimization (EHO-based parameter tuning)	88.7	87.5	86.9	87.1
Preprocessing + Optimization + Base ConGAN	92.4	91.0	90.6	90.8
Full Proposed Hybrid SEH-ConGAN	96.8	95.7	96.2	95.9

4.8 Discussion

Robot-assisted systems for animation production seek to automate labor-intensive processes, improve accuracy, and increase efficiency by combining robots with AI and deep learning approaches. In this context, sophisticated approaches have great potential but also significant limits. The EMOTION framework [16] allows humanoid robots to make socially suitable and expressive motions, which improves human-robot interaction; yet, it fails to catch delicate gestures and context-dependent signals, compromising naturalness in complicated circumstances. The E-FOOM with CBAM [17] increases character visual appeal and posture reconstruction, but it struggles with extreme postures, fast motions, and highly dynamic scenarios, potentially reducing animation accuracy. Similar to this, current motion generating models [24] (TSLA, ACTOR, MoGlow, VideoSwin) sometimes have poor trajectory smoothness, limited motion diversity, or decreased frame-level realism, which limits their capacity to produce completely varied and natural animations in robot-assisted production. The SEH-ConGAN improves animation production by effectively producing high-quality, diversified character movements with minimal manual labor. It also improves model parameters for greater accuracy and realism in animated scenes. The robot-assisted animation system performs well, with an accuracy of 0.96, a recall of 0.96, a precision of 0.97, and an F1-score of 0.96, indicating that automated motion creation is dependable, exact, and balanced for animation production.

5 Conclusion

The animation production process is usually labor-intensive, involving meticulous attention to character motion layout, scene composition, and post-production editing. It is made up of a motion capture interface, a robotic arm with feedback sensors, and a simulation environment for testing animation settings. Data preprocessing includes managing missing values and using Z-score normalization to provide consistent, high-quality AI model input. A sophisticated deep learning

algorithm. The SEH-ConGAN model is used to learn motion patterns, forecast human motion estimation, and transfer realistic motion styles. The AI models allow the system to acquire knowledge from animation data, forecast smooth character movements, and produce realistic animations depending on user input. The suggested method achieves better performance by using 5-fold cross-validation, and their average values are accuracy, recall, F1 score (0.96), and precision (0.97). A comparison of motion generation metrics reveals that SEH-ConGAN outperforms current models, obtaining the best MPJPE (16.7), FID (11.3), Smoothness (0.028), and Diversity (0.72), exhibiting higher animation realism, trajectory smoothness, and motion correctness. The robot-assisted animation system can struggle with highly creative activities, subtle character emotions, and complicated artistic styles, necessitating extensive setup and operator training. Future enhancements could involve AI-driven innovation, real-time human-robot cooperation, adaptive style learning, and support for 3D animation pipelines. Enhancing scalability and compatibility with cloud-based solutions can help to simplify production, decrease manual labor, and make animation processes more efficient and adaptable.

Competing interests

The authors have declared that no competing interests exist.

Data availibility statement

This study complies data availability policy. Data access arrangements align with the journal's guidelines and can be facilitated through the corresponding author.

Author Contributions

writing—original draft preparation:Hongping Tang
writing—review and editing:Hongping Tang
data curation:Hongping Tang

Reference

[1] Guo, Z. and Li, T., 2024. Practical analysis of virtual reality 3D modeling technology for animation majors based on predictive correction method. Informatica, 48(13). https://doi.org/10.31449/inf.v48i13.6129

[2] Ecole, L., Kim, W.T. and Yoon, J.S., 2023. Unity: A powerful tool for 3D computer animation production. Journal of the Korea Computer Graphics Society, 29(3), pp.45-57. https://doi.org/10.15701/kcgs.2023.29.3.45

[3] Yuanliang, W. and Zhe, Z., 2024. Integration effect of artificial intelligence and traditional animation creation technology. Journal of Intelligent Systems, 33(1), p.20230305. https://doi.org/10.1515/jisys-2023-0305

[4] Lungu-Stan, V.C. and Mocanu, I.G., 2024. 3D character animation and asset generation using deep learning. Applied Sciences, 14(16), p.7234. https://doi.org/10.3390/app14167234

[5] Tang, J., 2023. Graphic design of 3D animation scenes based on deep learning and information security technology. Journal of ICT Standardization, 11(3), pp.307-328. https://doi.org/10.13052/jicts2245-800X.1135

[6] Zhang, N., Meng, H., and Ju, M., 2024. Intelligent construction of animation scenes and dynamic optimization of character images by computer vision. Computer-Aided Design and Applications, pp.233-246. https://doi.org/10.14733/cadaps.2024.S25.233-248

[7] Hong, Z., Xu, X., and Liu, X., 2025. Application of virtual reality technology based on artificial intelligence in a 3D animated film storyboard. Discover Computing, 28(1), p.147. https://doi.org/10.1007/s10791-025-09670-7

[8] Nambiar, S., Wiberg, A. and Tarkian, M., 2023. Automation of an unstructured production environment by applying reinforcement learning. Frontiers in Manufacturing Technology, 3, p.1154263. https://doi.org/10.3389/fmtec.2023.1154263

[9] Zhao, J. and Zhao, X., 2022. Computer-aided graphic design for virtual reality-oriented 3D animation scenes. Computer-Aided Design and Applications, 19(1), pp.65-76. https://doi.org/10.14733/cadaps.2022.S5.65-76

[10] Ding, W. and Li, W., 2023. High speed and accuracy of animation 3D pose recognition based on an improved deep convolution neural network. Applied Sciences, 13(13), p.7566. https://doi.org/10.3390/app13137566

[11] Liu, X. and Zhao, H., 2025. MFFCN-GAN: Multi-scale feature fusion CNN with GAN for automated artistic scene generation in film animation. Informatica, 49(9). https://doi.org/10.31449/inf.v49i9.8903

[12] Liu, X., 2022. Animation special effects production method and art color research based on visual communication design. Scientific Programming, 2022(1), p.7835917. https://doi.org/10.1155/2022/7835917

[13] Wibowo, M.C., Nugroho, S., and Wibowo, A., 2024. The use of motion capture technology in 3D animation. International Journal of Computing and Digital Systems, 15(1), pp.975-987. http://dx.doi.org/10.12785/ijcds/150169

[14] Jiang, J. and Wang, X., 2024. Animation scene generation based on deep learning of CAD data. Computer-Aided Design and Applications, 21, pp.1-16. https://doi.org/10.14733/cadaps.2024.S19.1-16

[15] El-Raheb, K., Kougioumtzian, L., Kalampratsidou, V., Theodoropoulos, A., Kyriakoulakos, P. and Vosinakis, S., 2025. Sensing the inside out: An embodied perspective on digital animation through motion capture and wearables. Sensors, 25(7), p.2314. https://doi.org/10.3390/s25072314

[16] Huang, P., Hu, Y., Nechyporenko, N., Kim, D., Talbott, W., and Zhang, J., 2025. EMOTION: Expressive motion sequence generation for humanoid robots with in-context learning. IEEE Robotics and Automation Letters. https://doi.org/10.1109/LRA.2025.3575983

[17] Cao, W. and Huang, Z., 2025. Character generation and visual quality enhancement in animated films using deep learning. Scientific Reports, 15(1), p.23409. https://doi.org/10.1038/s41598-025-07442-3

[18] Zhang, N. and Pu, B., 2024. Film and television ani mation production technology based on expression t ransfer and virtual digital human. Scalable Computi ng: Practice and Experience, 25(6), pp.5560-5567. h ttps://doi.org/10.12694/scpe.v25i6.3351

[19] Racinskis, P., Arents, J. and Greitans, M., 2022. A motion capture and imitation learning-based approac h to robot control. Applied Sciences, 12(14), p.7186. https://doi.org/10.3390/app12147186

[20] Liu, Y., Li, L. and Lei, X., 2024. Automatic generati on of animation special effects based on computer vi sion algorithms. Computer-Aided Design and Appli cations, 21, pp.69-83. https://doi.org/10.14733/cada ps.2024.S23.69-83

[21] Wang, Z.S., Song, C.G., Lee, J., Kim, J.H., and Kim , S.J., 2022. Controllable swarm animation using de ep reinforcement learning with a rule-based action generator. IEEE Access, 10, pp.48472-48485. https: //doi.org/10.1109/ACCESS.2022.3172492

[22] Pibernik, J., Dolić, J., Mandić, L. and Kovač, V., 20 23. Mobile-application loading-animation design an d implementation optimization. Applied Sciences, 1 3(2), p.865. https://doi.org/10.3390/app13020865

[23] Kang, Y. and Kim, J., 2024. Animation character g eneration and optimization algorithm based on com puter aided design and virtual reality. Computer-Ai ded Design and Applications, 21(S14), pp.46-62. ht tps://doi.org/10.14733/cadaps.2024.S14.46-62

[24] Li, Q., Sun, T., and Zhang, M., 2025. Deep learning -driven animation: Enhancing real-time character m otion synthesis. IEEE Access. https://doi.org/10.11 09/ACCESS.2025.3623285

Hybrid Panel Data Forecasting for Coastal Flood Hazard Assessment: A Spatial Data Mining Approach Using Exponential Smoothing and Neural Networks

Dimara Kusuma Hakim[1*], Rahmat Gernowo [2], Anang Widhi Nirwansyah [3]
[1]Doctoral Program of Information Systems, Diponegoro University, Indonesia
[2]Department of Physic Faculty of Mathematics and Natural Sciences, Diponegoro University, Indonesia
[3]Department of Geography Education, Universitas Muhammadiyah Purwokerto, 53182, Indonesia
E-mail: dimarakusumahakim@gmail.com
[*]Corresponding author

Keywords: flood hazard, panel data, forecast, coastal, exponential smoothing, neural network

Received: Januar 30, 2025

Due to climate change–induced sea level rise, coastal flood risk is increasing significantly, creating an urgent need for improved flood risk assessment and mitigation strategies. Effective flood analysis requires both time series forecasting and hazard risk evaluation, including the determination of appropriate weights for flood-related parameters such as elevation, runoff, distance to the shoreline, and sea level. Traditional forecasting methods like exponential smoothing are limited in capturing relationships between variables, while neural networks can model non-linear interactions but are less commonly applied to long-term forecasting. To address these limitations, this study proposes a hybrid method that integrates Exponential Smoothing (ES) and Neural Networks (NN) for panel data analysis, where ES identifies trends and seasonal patterns and its output is used as additional input for NN. The NN is also employed to objectively determine parameter weights, reducing the subjectivity of conventional AHP-based approaches. The method integrates geographic and climatic variables, including wind speed, temperature, sea surface pressure, and rainfall, and is applied to coastal areas in Semarang, Demak, and Jepara, Indonesia. Results show that the hybrid model outperforms standard ES and NN methods, achieving flood forecasting errors (MAPE) between 3.5% and 8.3% and parameter weighting accuracy of 88–94%, contributing to a more holistic and reliable flood risk analysis.

Povzetek: Raziskava predlaga hibridni pristop, ki združuje eksponentno glajenje in nevronske mreže za natančnejše napovedovanje obalnih poplav ter objektivno določanje uteži poplavnih dejavnikov, s čimer izboljša oceno poplavnega tveganja na podlagi geografskih in podnebnih podatkov.

1 Introduction

Rob floods are one type of flood that increasingly threaten coastal areas worldwide, occurring as a phenomenon when high tide sea water overflows onto land. Global warming causes rising sea levels, increase the risk of coastal flooding by up to five times this century worldwide, putting more than 70 million people in danger in increasingly expanding floodplain areas, with projections of loss of land and critical infrastructure due to permanent inundation [1]. The fact that many major cities and communities are situated along the shore, making more people and infrastructure susceptible to coastal flooding, makes this situation much worse. By 2050, hundreds of densely populated coastal cities are expected to face increased flood risk, with that risk doubling by 2100 [1]. This increase could lead to a setback in human development globally, especially in coastal areas that are centers of important social and economic activity. In addition to being caused by human activities such as urbanization and land-use change, climate change factors also exacerbate these conditions through rising sea levels and increasing frequency of floods. Therefore, a deep understanding of flood risk and developing effective mitigation strategies are critical to protecting coastal communities from such negative impacts.

Hazard risk analysis and time-series forecasting are important approaches in flood mitigation [2], [3], [4]. This involves assessing the flood potential in an area, with: evaluation of historical flood data [5], [6], [7], [8], weight determination [2], [9], [10], and future climate projections to predict where and when floods might occur [11], [12]. Predictive models in this field are increasingly being enhanced by the application of Spatial Data Mining (SDM), which enables the analysis of large datasets to identify trends and improve forecast accuracy [13], [14]. SDM plays a crucial role in enhancing predictive models for flood forecasting by enabling the analysis of large datasets to identify trends and improve forecast accuracy. It integrates various data types, including historical flood data and geographical,

climatic, and land use variations, essential for practical flood risk analysis. The application of SDM allows for identifying non-linear correlations between variables that traditional forecasting methods, such as univariate approaches, often overlook. By combining SDM with techniques like Artificial Neural Networks (ANN) and Exponential Smoothing (ES), predictive models can achieve significantly improved accuracy, as evidenced by a hybrid model that achieves an error smaller than single methods [1], [2], [6]. This methodological advancement is significant in regions like Indonesia, where geographical and climatic variances can significantly affect flood dynamics.

Long-term forecasting techniques used in predicting floods in time series generally employ a univariate approach [15], with time series data, such as the exponential smoothing method, which has the advantage of simplicity in the mathematical formulas used. However, exponential smoothing has the disadvantage of being limited in depicting relationships with other influencing variables [16], [17]. The solution to the problem of the relationship between variables causing floods can use non-linear approach neural networks, with cross-sectional data, which has the advantage of making fairly accurate estimates in the short term but does not have the advantage of long-term forecasting like the univariate approach [18]. Both types of data, time-series and cross-section data, when combined in the form of panel data, can be used for time-series forecasting analysis that links the relationships between variables in making forecasts, by combining both approaches (time-series forecast and non-linear estimation) with hybrid method (Triple Exponential Smoothing and Neural Network). The hybrid method approach is expected to offer solutions that not only improve the accuracy of predictions but also adaptive flexibility to environmental dynamics, making them a relevant and strategic topic in today's era of climate change [13], [19], [20], [21].

Previous studies with various methods (ARIMA, SARIMA, LSTM, etc.) have highlighted the effectiveness of different forecasting models in coastal flood forecasting but have not consistently reported MAPE values as part of their evaluation metrics, focusing only on RMSE and MAE [22], [23], [24]. Although Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are valuable tools for assessing model performance in a particular context or study, they have limitations in terms of cross-study comparisons due to their subjectivity related to the characteristics of the data [25], [26]. In contrast, MAPE provides a more standardized approach that facilitates comparisons across multiple scenarios [27], [28], so the use of MAPE in coastal flood forecasting studies is a novelty.

To provide accurate visualization in the form of a flood hazard map, it is critical to determine the proper weight of flood risk parameters so that each pixel on the map can be given a representative risk score. This weight reflects the degree of influence of each variable, such as elevation, runoff, distance to water sources, and sea level, on the potential for flooding in the location. Weighting is essential because of geographic variation. In flood hazard risk analysis, it is necessary to determine the weight values for various parameters that influence flood risk. Geographic, climatic, and land use variations in different regions of Indonesia cause the flood hazard weights to differ in each area. The Analytic Hierarchy Process (AHP) method, which is often used to determine parameter weights, is quite effective in empirical analysis [29], [30]. However, this method needs to be improved due to subjectivity and the potential bias of experts [31]. AHP relies heavily on subjective judgments for assigning weights to different criteria, which can introduce bias and inconsistency [32], [33]. Additionally, AHP may struggle with handling complex, nonlinear relationships among flood risk factors and can be limited in capturing the dynamic interactions between variables over time and space.

In this study, we propose an innovative solution using Historical Flood Data-Based Neural Networks to determine parameter weights more objectively and accurately. This approach is expected to reduce bias and enhance the reliability of flood risk analysis through data-based optimization [2], [33]. Neural network-based approach addresses these limitations by learning directly from data, which allows it to model complex, nonlinear patterns in flood risk factors without requiring subjective weight assignments. Neural networks can integrate multiple influential variables and their interactions more effectively, improving prediction accuracy and robustness. Moreover, when combined with GIS and remote sensing data, neural networks provide a more data-driven, adaptive, and precise flood susceptibility mapping compared to the more rigid and expert-opinion-dependent AHP.

The objective of this study is to evaluate the effectiveness of combining Neural Networks and Exponential Smoothing in improving long-term and short-term flood forecasting accuracy in coastal areas. We hypothesize that combining Neural Networks and Exponential Smoothing improves long-term and short-term flood forecasting accuracy compared to single-model approaches. The study focuses on Semarang, Demak, and Jepara due to their significant exposure to flooding events, as evidenced by historical data indicating a high frequency of flood occurrences in these areas. Additionally, these locations are characterized by unique environmental conditions that make them particularly susceptible to the impacts of climate change, thus providing a relevant context for this research.

2 Data

The study was carried out in coastal areas representing tropical climates in Indonesia, specifically in Semarang City, Demak, and Jepara, covering an area of approximately 2,291.456 km² with a total population of 4,819,953 people, where Semarang City has 373.78 km² and 1,694,743 people, Demak has 897.43 km² and 1,252,970 people, and Jepara has 1,020.246 km² and 1,872,240 people based on the latest measurements [34]. The three areas were chosen because they are prone to

hydrometeorological disasters on the north coast of Java, especially coastal floods, caused by a combination of high rainfall, extreme sea tides, and land subsidence due to both human activities and natural processes, and they also have a long history of flood events [35], [36], [37].

For the analysis of coastal flood hazards, several variables are used: sea level data, elevation, runoff (from land cover), and distance to the coastline. Elevation, runoff, and distance are variables that tend to remain constant (not experiencing significant changes), whereas sea level data is a variable that varies over time.

In contrast to other variables, sea level is changing continuously over time as a water surface along the coast. Tidal gauge records and sea tide models are used to recreate normal high tide conditions as well as infrequent but significant extreme events [38]. The study makes use of historical sea level data for the years 1995–2024 from the Indonesian Geospatial Reference System (SRGI) [45]. Many time-based weather variables, such as U10 & V10 (10m Wind Speed), D2M (2m Dew Point Temperature), T2M (2m Temperature), MSL (Mean Sea Level Pressure), SP (Surface Pressure), TP (Total Precipitation), and SWH (Significant Wave Height), are used to forecast sea level with the assumption that all of these variables have an impact on sea level [47], [48], [49]. Wind speed has an impact on water flow and can lead to storm surges, which raises the possibility of coastal flooding [39], [40]. Rainfall patterns and cloud formation are influenced by atmospheric humidity and condensation, which are influenced by the dew point and air temperature [40]. In addition, surface pressure and average sea surface pressure have a direct impact on ocean and atmospheric dynamics that affect water level variations [41], [42]. Both the intensity of tropical storms and the thermal expansion of seawater are influenced by sea surface temperature [42]. Meanwhile, total precipitation increases the amount of water on land [42], which can make runoff worse. Information about wave patterns that can either increase or lessen the impact of extreme tides in coastal areas can be found in the direction of the waves [42]. By combining these variables, the prediction model is able to account for the intricacy of the atmosphere-ocean interaction, producing estimates of flood risk and sea level rise that are more accurate. Hourly weather data recorded as time series with a resolution of roughly 27.83 meters is the source of ERA5 hourly data on single levels (Copernicus), from which weather data is derived [46].

Elevation is fundamental in flood threat analysis because it determines which areas are physically vulnerable to inundation. For example, low-lying areas are more susceptible to flooding due to their proximity to sea level, directly affecting the level and depth of potential flood inundation [38]. Elevation data is obtained from DEMNAS Indonesia (National Digital Elevation Model), provided by the Geospatial Information Agency (BIG).

Runoff values, derived from land cover data conversion, represent how much rainfall becomes surface runoff rather than infiltrating into the ground [38]. Therefore, areas with impermeable surfaces or specific land uses generate higher runoff, increasing the flood threat. Runoff data is derived from land cover: Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3 [43], which is then converted based on the Rational Method Runoff Coefficient [44].

Distance to the coastline is essential because closeness increases exposure: locations closer to the coast are at greater risk from storm surges and sea level rise events. A spatial relationship between the location of each pixel and its distance from the coastline is concluded when creating flood threat maps [38]. The distance to the coastline is essential because proximity increases exposure: locations closer to the coast are at greater risk from storm surges and sea level rise events, concluding that there is a spatial relationship between the location of each pixel and its distance from the coastline when creating flood threat maps. Distance to the coastline is calculated by measuring the distance of each pixel from Shapefile to the nearest coastline in each area.

3 Method

3.1 State of the art and related works

Forecasting, especially related to flooding, is a research field that continues to be discussed, with various approaches being used. One of the main focuses is on coastal locations vulnerable to rising sea levels and land subsidence, which significantly increases the risk of flooding [52] and ultimately negatively impacts the local economy [53]. Some topics related to Coastal Flood Forecasting (Fig. 1) are flood risk and adaptation response modeling, hydrodynamic modeling, and flood risk assessments.

Coastal flood risk modeling and adaptation response assesses future flood risk and formulates effective adaptation strategies. By analyzing climate change scenarios and their consequences on flood risk, this method integrates physical and social models to understand potential impacts comprehensively [54].

Hydrodynamic modeling examines and predicts water flow behavior and its interactions with the environment [55]. Using principles of physics, mathematics, and engineering, these models accurately simulate water movement under various conditions, including floods, waves, currents, and other hydrologic phenomena.

Flood risk assessment systematically identifies, analyzes, and evaluates the risks associated with flooding [56], [57], [58]. This approach seeks to understand the potential hazards of flooding, evaluate their likely impacts, and assess the vulnerability of infrastructure and populations to flood events.

These approaches can stand alone or be enhanced through integration with statistical or data mining techniques [56], strengthening the ability to predict and respond to coastal flood hazards. This study seeks to explore flood risk assessment through an innovative approach that leverages data mining for accurate estimates.

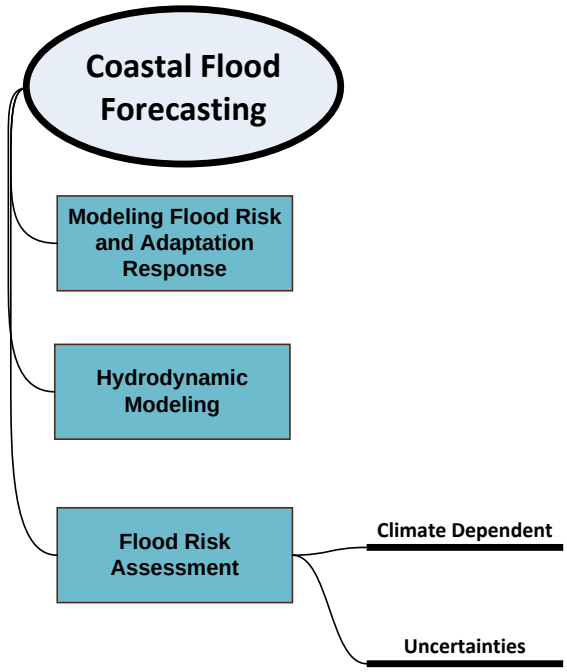


Figure 1: Coastal flood forecasting mind map

Several methods have been used in sea level forecasting associated with risk assessment (Table 1), with various advantages and disadvantages. Using a single method in prediction has the advantages of simplicity and clarity. However, this method has several limitations: limitations in capturing data complexity, risk of overfitting or underfitting, and difficulty handling data variability. In terms of data complexity, forecasting methods are generally univariate (e.g., ARIMA, ES, etc.) which do not consider the influence between variables as in regression approaches (linear regression, neural networks, etc.), so a hybrid approach is proposed by combining forecasting and the influence between variables.

Table 1: Sea level forecasting related works

Data	Techniques	Accuracy Metrics	Limitations
Tide Gauge Data	ARIMA [59], [60], [61]	RMSE	Limited to linear relationships, struggles with non-stationary data
Satellite Altimetry	Neural Network Auto-Regressive (NNAR) [62], [63]	MSE, MAE, RMSE	Requires large datasets, computationally intensive
Tide Gauge + Satellite Data	Extreme Gradient Boosting (XGBoost)	MSE, MAE, RMSE	Sensitive to outliers, overfitting risk
Tide Gauge Data	Facebook Prophet (FProphet) [64]	MSE, MAE, RMSE	Assumes additive seasonality, less effective for multiplicative trends
Tide Gauge Data	Exponential Smoothing [65]	RMSE	linear trends, sensitive to sudden changes
Tide Gauge + Satellite Data	Long Short-Term Memory (LSTM)	MSE, MAE, RMSE	Requires large datasets, computationally intensive
Tide Gauge Data	Moving Average	RMSE	Limited to short-term forecasting, less effective for capturing seasonality
Tide Gauge Data	ARMA	RMSE	Limited to linear relationships, struggles with non-stationary data
Tide Gauge + Satellite Data	Recurrent Neural Network (RNN)	MSE, MAE, RMSE	Requires large datasets, computationally intensive

In several recent studies, hybrid approaches can improve prediction accuracy by reducing bias and variance, thus handling non-linear relationships and complex interactions among environmental variables [66], as well as capturing temporal patterns and long-term dependencies in data [67], which are very important for hydrological predictions. This approach adds an essential dimension to flood vulnerability analysis by emphasizing the use of predictive machine learning to assess relationships among variables [68], making it possible to predict spatial vulnerability levels in areas.

Exponential Smoothing effectively captures trends and seasonality in time series data, providing stable and interpretable predictions. However, this method struggles with non-linear patterns and complex interactions between variables. On the other hand, neural networks excel at modeling non-linear relationships. They can learn complex patterns from large datasets, but they often operate as a black box, making interpretation difficult. Integration of Exponential Smoothing and Neural Networks by combining the strengths of both methodologies is expected to improve the accuracy of sea level forecasting for coastal flood hazard prediction [69].

Integrating these two methods is expected to leverage the stability and interpretability of Exponential Smoothing while leveraging the predictive power of Neural Networks, ultimately resulting in more robust and reliable flood hazard forecasts that can adapt to changing environmental conditions.

3.2 Flood hazard in coastal area

In several disaster risk studies, especially the flood hazard assessment conducted in this study, hazards are measured using a multi-criteria approach, where several variables that make up hazards are given scores and weights according to the level of impact, such as in Indonesia, which uses the Indonesia Disaster Risk Index [70], [71]. The formula used is the classic weighted sum, the sum of the multiplication results between the weight and the value/score of each variable. It is commonly used in risk analysis, multi-criteria decision-making, and hazard evaluation (Eq.1). The Variable Score (S_i) is the value of the intensity of the hazard based on various

factors. Weights (w_i) can be determined in various ways: expert judgment, AHP, or data-driven (used in this study) to assess the contribution of each factor to the overall hazard.

$$\text{Hazard Index} = \sum_i^n w_i S_i \quad (1)$$

$$\text{H} = w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4 \quad (2)$$

The Hazard Index (H) assesses an area's danger level based on several factors (Eq.2). The first factor, elevation (S_1), is the height of a location from sea level that affects the risk of flooding [71]. The second factor, runoff (S_2), is the water flow above the surface that exceeds the soil's capacity to absorb water [72]. The third factor, distance to the coastline (S_3), is the distance of a location from the coastline, where locations closer to the coast may be more vulnerable to hazards such as tsunamis and hurricanes [73]. The last factor, sea level (S_4), measures the influence of tides that can affect flood risk in coastal areas [74]. Each of these factors is weighted (w_1, w_2, w_3, w_4) according to the degree of its influence on the overall Hazard Index (H).

All variables are stored in raster format (GeoTIFF), and calculations are performed on each pixel with a resolution of about 8 meters (adjusting to the initial Raster: DEM). Each raster has been cut and re-sampled to ensure consistency and accuracy of the data.

3.3 Neural network data driven weighting

The use of feature importance in determining flood hazard weights provides a more objective and data-driven approach compared to traditional methods such as the Analytic Hierarchy Process (AHP), which heavily relies on the subjective judgments of experts. With feature importance, the relative contribution of various parameters is assessed. Differences between regions significantly affect the results of flood hazard assessments because each area has unique geographic characteristics [75], [76], climate, land use, and socio-economic conditions, all of which influence how the flood risk parameters play a role.

Weighting using neural networks can be done by calculating feature importance and measuring the variance of the weights connected to each input feature during the training phase [77], [78]. This process involves assessing how much each feature affects the output by observing the changes in related weights. The more significant the weight change, the more important the corresponding feature influences the output [77]. This method is beneficial for understanding which input variables impact the model's predictions most, thereby allowing for better interpretation and optimization of model performance. The final feature importance scores are calculated based on the variance of the weights combined with the final weights for each input feature; this helps identify the most critical factors affecting the model's output, such as flood hazard assessment.

Several feature variables are used to calculate feature importance in the flood depth prediction model: elevation, runoff, distance to the coastline, and sea level [79], [80]. The flood depth class variable plays a role as the target to be predicted. The value of the flood depth class is obtained from the flood depth classification process, which is measured on a meter scale and then categorized into certain classes to facilitate analysis and modeling.

Each of these feature variables has an important role in flood modeling. Elevation determines the height of the land relative to sea level, which directly affects the risk and level of flooding [81]. Runoff describes the flow of surface water from rainfall and runoff, contributing to the water volume that can cause flooding [82]. Distance to the coastline affects the potential impact of coastal flooding because the closer a location is to the coast [82], the more likely it is to be flooded due to rising sea levels. Sea level is in the form of sea level fluctuations that can worsen flood conditions, especially in coastal areas [82].

Calculating feature importance in neural network [77] starts with data preparation, building a model, training, evaluation, and the last is feature importance analysis (Fig. 2). Building model involves: determining the number of hidden layers and neurons, using ReLU (Eq.3) and softmax (Eq.4) activation functions, and finally initializing weights and biases.

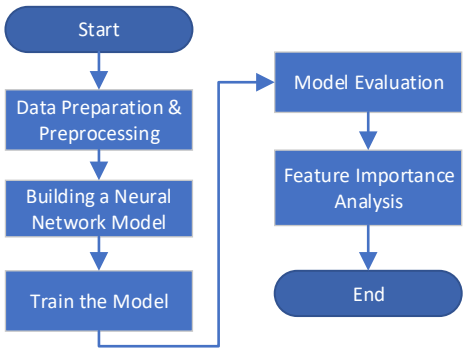


Figure 2: Weighting with NN

$$f(x) = \max(0, x) \quad (3)$$

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (4)$$

The ReLU (Rectified Linear Unit) activation function is a straightforward function in neural networks, which converts all negative input values to zero and maintains positive values as they are (Eq.3), with a graph in the form of a straight line on the positive side and a flat line on zero for the negative side, so that non-linear relationships are captured yet efficiently while keeping computation light [83]. The softmax activation function (Eq.4) is used on the output layer of the multiclass classification model to convert the raw output value (Z_i) into an interpretable probability (P_i) by converting the input vector ($z = (z_1, z_2, \ldots, z_K)$ into a probability distribution ($p\ [p_1, p_2, \ldots\ldots, p_K]$), each output element is in the range of 0 to 1 and the sum of all components is 1, so that softmax provides a prediction in the form of the odds

of each class (the class with the highest probability is selected as the prediction result) [83].

During the training phase (Fig.2), backpropagation is performed to calculate the output, calculate MSE & MSE, and update the weights based on the loss gradient with the Adam optimizer (Adaptive Moment Estimation) to adaptively adjust the learning level for each parameter to accelerating convergence and improving model performance [84]. Adam is an optimization algorithm that combines the concepts of momentum and adaptive learning rate concepts by calculating the moving average of the first gradient (momentum) and the second quadratic gradient to adjust parameter updates adaptively. At iteration t, Adam calculates the biased first moment estimate m_t (Eq. 5), which is a moving average of current gradients gt weighted by decay rate β_1; similarly, it computes the biased second raw moment estimate v_t (Eq. 6), a moving average of squared gradients weighted by decay rate β_2. These estimates are then bias-corrected to obtain unbiased moments (Eq. 7 and Eq. 8). Finally, parameters are updated (Eq. 9) where $\alpha > 0$ is the learning rate controlling step size, and small constant ($\epsilon > 0$) ensures numerical stability to avoid division by zero. This mechanism allows Adam to efficiently adapt to complex loss landscapes, accelerating convergence during training.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\mathrm{g_t} \quad (5)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\mathrm{g}_t^2 \quad (6)$$

$$\widehat{m_t} = \frac{m_t}{1 - \beta_1^t} \quad (7)$$

$$\widehat{v_t} = \frac{v_t}{1 - \beta_2^t} \quad (8)$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon} \quad (9)$$

The Magnitude of Weights (Absolute Weights) on the Input Layer is used to calculate feature importance, which is often used to measure feature importance in neural networks. It is done by referring to the absolute value of the weights (w_{ij}) that connect the input neuron to the neuron in the next layer (Eq. 10), where the i index indicates the i input feature and the j index indicates the neuron in the hidden layer. By calculating the sum or average of the absolute value of the weight, the contribution value (influence) of the feature on the output of the model is obtained. The greater the magnitude of a feature's weight, the more important the role it plays in the model's prediction process because small changes to that feature will significantly impact subsequent neuron activation.

$$\mathrm{FI_i} = \sum_{\mathrm{j=1}}^{\mathrm{n}} |W_{ij}| \quad (10)$$

3.4 Time series, cross-section, and panel data

Time series data is a sequence of data points collected or recorded at consecutive time points (e.g., Table 2), with the characteristics of a date (period) and one actual variable [85] , which are analyzed using a univariate time series forecasting approach such as the Triple Exponential Smoothing (TES) method [17].

Table 2: Sample of time series data

Date	Sea level (cm)
2022-11	54.3
2022-12	55.1
2023-01	49.8
2023-02	51.3

Cross-sectional data is data collected by capturing snapshots of various variables without considering changes over time. Cross-sectional data (example: Table 3) are analyzed using a non-linear estimation approach [86], [87] like Neural Network (NN), which is effective in capturing complex and non-linear relationships in the data. NN utilizes learning algorithms to model intricate patterns and interactions among variables, making it suitable for handling cross-sectional data.

Table 3: Sample of Cross-sectional data

Location	Temperature	Wind Speed	Sea Level
Demak	30.82	1.1302	54.3
Demak	30.47	1.4842	55.1
Jepara	29.87	0.8074	49.8
Jepara	30.48	1.2403	51.3
Semarang	29.54	1.1908	44.7
Semarang	30.45	1.6641	55.6

Panel Data, which is a combination of time series and cross-sectional data, for example, sea level data collected from several locations at different points in time (example: Table 4).

Table 4: Sample of panel data

Date	Location	Temperature	Wind Speed	Tide
2022-12	Demak	30.82	1.1302	54.3
2023-01	Demak	30.47	1.4842	55.1
2022-12	Jepara	29.87	0.8074	49.8
2023-01	Jepara	30.48	1.2403	51.3
2022-12	Semarang	29.54	1.1908	44.7
2023-01	Semarang	30.45	1.6641	55.6

The data panel has a key advantage over purely time-series or cross-sectional data by integrating diverse data sources from time and space in flood risk assessments, emphasizing the need for dynamic models that can incorporate changing variables such as land use, climate, etc [88]. The evolution of flood modeling techniques increasingly prioritizes integrating spatial and temporal

data, supporting models that can dynamically respond to changes in environmental and anthropogenic factors [2]. Various case studies and methodologies also highlight the importance of applying advanced data techniques, including spatial-temporal modeling, to reduce uncertainty in flood risk assessments more accurately and effectively [89]. Panel data enables evaluation how flood risk mitigation measures perform over time and across regions.

In this study, UTC was used as a global time standard, which provides a consistent reference, which is officially abbreviated from Coordinated Universal Time as a neutral compromise between the English terms (Coordinated Universal Time) and the French term (Temps Universel Coordonné)—used [90]. Panel data is stored in Excel format and then imported into Pandas DataFrame using Python to facilitate data analysis and processing.

3.5 TES & NN combination to forecast panel data

Temporal dependencies in the panel data are included in two main steps [91] that carried out on the panel data; the first step focuses on the classic statistical method of Triple Exponential Smoothing (TES) to capture trend and seasonal patterns in the data, while the second step combines the approach with a NN model that utilizes feature selection [92] in the form of the selection of relevant weather features as inputs to capture complex non-linear relationships. The combination of the results of the two and aims to produce predictions that utilize each method's strengths and optimize input features [93], which models non-linear patterns and complex interactions between variables. Thus, this integration allows the model to effectively accommodate temporal dependencies while improving the accuracy of predictions on multivariate panel data.

The use of the TES for time series data is characterized by calculations: Level, Trend and Seasonal [94], which is a moving forecasting method that gives weight in stages to the latest data, so that if the data is new, the forecasting value will be updated.

The data is divided into a period (annual / biennial / five years / etc.), the initial level value is obtained from the average actual value in the first period (Fig.2), the initial trend value is obtained from the average value of the difference in the actual value of each month in the same 2 periods and divided by the number of months, and the initial seasonal value for each month is the actual value divided by the Initial Level value. Then the Level (L_t), Trend (T_t), Seasonal (S_t) and Forecast (F_{t+m}) are calculated. Smoothing parameters used: α as level parameters, β as trend parameters, and γ as seasonal parameters. Y_t is the value of observation at time (t). S_{t-m} is a seasonal factor in the previous period, L_t is the smoothed level, T_t is a smoothed trend, and (m) is the length of the season.

$$L_t = \alpha\left(\frac{Y_t}{S_{t-m}}\right) + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (6)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (7)$$

$$S_t = \gamma\left(\frac{Y_t}{L_t}\right) + (1-\gamma)S_{t-m} \quad (8)$$

$$F_{t+m} = (L_t + mT_t)S_{t-m+k} \quad (9)$$

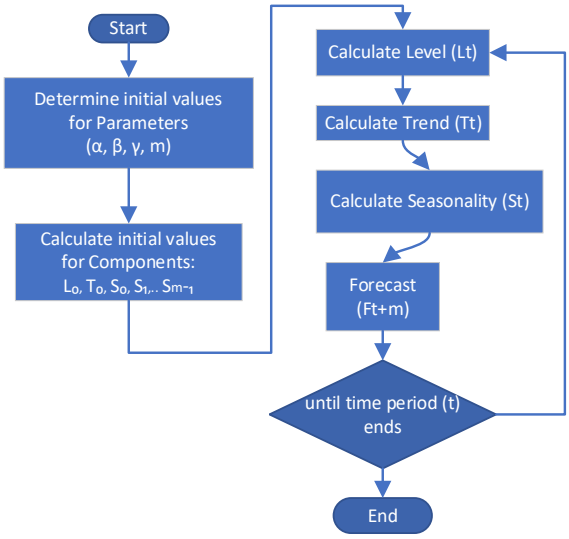


Figure 3: TES

The calculation of some of these parameters (Fig.3) is repeated until a minimum error is obtained through the optimization process up to the specified period (t). Each iteration is performed by updating the model parameters using the L-BFGS-B algorithm that efficiently searches for the optimal solution and ends when the convergence criteria are met (a slight change in the value of the error function or reaching a predetermined maximum number of iterations). The resulting model is expected to have optimal performance in predicting data in a certain period. Various optimization methods have been tried to perform smoothing parameter optimization, including popular algorithms such as Adam, Stochastic Gradient Descent (SGD), etc. However, the results of the experiments showed that L-BFGS-B provided the highest accuracy compared to these alternatives. L-BFGS-B consistently produces more optimal objective function values and predictions with the highest precision [95], so it is considered superior in smoothing parameter optimization. L-BFGS-B uses Hessian matrix approximation efficiently, allowing faster and more stable convergence than gradient-based methods such as SGD or Adam [84].

To make predictions from the forecast results (cross-sectional data), NN is used on weather data to predict sea level (Fig.4). The data is divided into training and testing data, each covering a specific year period. The features used to train the model are several columns that have been normalized. Then, several hidden layers, neurons, and the ReLU activation function (Eq.10) are used. The prediction results are evaluated using the MAPE (Eq.11) and MSE (Mean Squared Error) (Eq.12) metrics [94].

$$\mathrm{ReLU}(x) = \max(0, x) \quad (10)$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \tag{11}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{12}$$

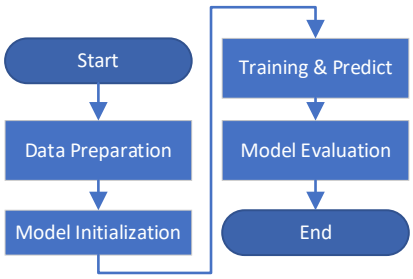


Figure 4: NN for prediction

There is no set rule for choosing the number of hidden layers or neurons. This decision is highly dependent on the data's nature and the modeling's purpose, so it is often trial-and-error or validation-based tuning [96], [97]. The NN architecture design chosen in this research, which is six hidden layers with 50 neurons each, is based on empirical experiments and the complexity of the modeled data. The architecture was chosen to balance the capacity of the model to study complex patterns while avoiding over-installing. Using multiple hidden layers allows the model to capture hierarchical features in the data, which is crucial in flood prediction by considering the relationships between variables.

The activation functions, ReLU, were selected based on performance in capturing non-linear relationships [98]. ReLU was chosen for its ability to reduce the problem of disappearing gradients [99], allowing for faster convergence during training [100], [101].

The hybrid forecasting method that combines TES and NN aims to solve panel data problems, improving prediction accuracy by leveraging the strengths of each technique, resulting in forecasting data that considers the influence between variables.

The first step is to forecast each variable using TES, effectively capturing seasonal patterns and trends in the data. The forecast results from TES are combined into a single dataset used as input for the NN model. NN is trained using this dataset to capture non-linear patterns and complex interactions between variables that TES may not capture. After the NN model is trained, predictions are made on the testing data, and the results are evaluated using metrics such as MAPE and MSE to measure the model's accuracy. By combining these two methods, the forecasting results are expected to be better than the univariate approach, as TES handles seasonal patterns and trends while Neural Networks capture complex prediction patterns.

3.6 TES & NN combination to forecast panel data with predicted bias

Bias refers to systematic errors that occur during training, leading to an inaccurate representation of the studied population [102], [103]. To improve prediction accuracy, the value of bias is optimized during the training process. This optimized bias can then be used in ensemble methods, combining predictions from multiple models to enhance overall accuracy [104]. Specifically, the output from neural networks, including the optimized bias, can be combined with the output from other models to generate final predictions. This approach aims to correct bias in time series data and improve the accuracy of forecasting models.

Analyzing the panel data with Average Bias involves two key stages: TES is used to identify trends and seasonal patterns within the dataset; subsequently, this method is integrated with the NN model that employs feature selection by choosing pertinent weather variables as inputs to model complex non-linear relationships. By combining the outputs of both processes and applying bias correction, the approach aims to generate predictions that leverage the strengths of each technique while optimizing the selected input features.

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{y_i} - \hat{y}_\mathrm{i}) \tag{13}$$

Bias in data mining refers to systematic errors that occur during training, leading to an inaccurate representation of the population being studied. Bias can arise from the data used or from the algorithm itself that shows a certain tendency. Bias assessment can be done data-driven using an estimation approach, such as the NN method. In a NN, bias is a parameter that is added to the input of each neuron before an activation function is applied. The Bias prediction carried out in this study is post-processing by adjusting the model output to be fairer or correcting the prediction based on specific metrics [105]. In detail, it is calculated by predicting the bias value for each sample in the test data using a trained model and then taking the average of all these predictions ((Eq.13)) to obtain one representative value [106]. This average value is then normalized by dividing it by the number of features (variables) used in the model. The result is a scalar value representing the overall bias estimate after normalization. The use of a single bias value from a NN (or similar method) can be used in ensemble methods, which combine predictions from multiple models to improve accuracy. In this case, the output from the NN (including the bias that has been optimized during training) can be combined with the output from other models to make the final prediction.

3.7 Forecasting model accuracy

To evaluate how well a model can predict the true values, including in the context of flood forecasting (sea level rise), several metrics are often used: MSE, RMSE and MAPE.

MSE (Mean Squared Error) measures the average of the squared errors between the predicted and actual values. While RMSE (Root Mean Square Error) measures the average of the squared errors between the predicted and actual values, then takes the square root. MSE and RMSE give a larger penalty for larger errors and are the basis for calculating RMSE [107]. MSE is the basis for calculating RMSE which is more sensitive to outliers [13], [108] so it is important in flood forecasting because flood events often involve extreme values (severe flood events).

MAPE (Mean Absolute Percentage Error) measures the relative error between predicted and actual values in percentage form. Unlike other metrics, MAPE is objective to the data and easy to interpret, as it is expressed as a percentage. It can be classified into the following categories [107]: MAPE < 10% is considered very good, 10% ≤ MAPE < 20% is good, 20% ≤ MAPE < 50% is sufficient, and MAPE ≥ 50% is deemed poor. All metrics are used, as an approach to evaluate how well the model can predict flood peaks or flood occurrence times.

A hybrid method between NN and TES is used, aiming to combine the advantages of both techniques to improve forecasting accuracy. Then, the gradient descent method is used as an optimization to obtain optimal weight values for NN and alpha (level), beta (trend), and gamma (seasonal) values for TES. The result of the analysis of the gathered data will be a geographic information system that shows maps of Flood Hazards (H).

3.8 Raster visualization for flood hazard analysis

Raster visualization for flood hazard analysis plays a crucial role in disaster risk management by providing a detailed, spatially explicit representation of flood hazard levels across a region [109]. Flooding poses significant threats to lives, infrastructure, and economies, especially in coastal and low-lying areas [110]. Accurate flood hazard maps enable policymakers, urban planners, and emergency responders to identify vulnerable zones, prioritize mitigation efforts, and design effective evacuation plans, ultimately reducing the impact of flood events on communities.This process begins with creating a raster for elevation by combining and cropping several GeoTIFF files from the Digital Elevation Model (DEMNAS) according to the region's boundaries [111], [112]. Using raster-based methods for final visualization allows for a fine-grained depiction of flood hazard levels, where each cell is assigned a score based on predicted or observed values, effectively highlighting areas with varying flood hazard [110], [113], [114].

The coordinate reference system (CRS) is standardized to EPSG:4326, and the spatial resolution is set to 8.35 meters, matching the original resolution of the DEMNAS data [115]. Maintaining this high resolution preserves critical topographic details for accurately analyzing elevation variations—a key determinant of flood risk. Furthermore, consistent resolution facilitates the integration of diverse datasets, enhancing the reliability and informativeness of the resulting flood hazard map.

Elevation data undergoes min-max normalization, where higher elevations correspond to lower flood risk scores, reflecting the inverse relationship between elevation and flood susceptibility. Land cover data is converted into runoff values using the Rational Method Runoff Coefficient, contributing to the overall flood hazard score by accounting for surface permeability and water retention capacity. Distance to the coastline is calculated using a shapefile from the Indonesian Landmark, with an Euclidean distance transform applied to measure proximity to the sea. This distance is normalized and inverted, as areas closer to the coastline generally face higher flood risks due to storm surges and sea-level rise.

A sea level raster is created by combining sea level scores with distance scores linked to estimated sea level values. The method produces a comprehensive flood hazard map by integrating all these factors—elevation, land cover runoff, coastline proximity, and sea level—into a weighted composite raster. This nuanced visualization supports targeted flood risk assessment and informs sustainable coastal management strategies, helping to safeguard vulnerable populations and infrastructure from future flood events.

4 Result

The analysis begins by determining the weight of each coastal flood hazard factor, with each factor assigned a weight based on its level of influence on the coastal flood risk. After the weights of the hazard factors are determined, the next step is to perform univariate forecasting for each variable using the TES. The results of this univariate forecasting are then used as input for the next prediction step using NN. This prediction aims to determine the flood depth class while considering the complex interactions between variables.

4.1 Weight determination with NN

Determining flood hazard weights using neural networks involves several steps, from data collection and processing to model training and evaluation. The first step is to determine the coastal flood database, which contains a history of flood and non-flood events. This data includes several parameters that occur during floods (average height, average runoff, average closest distance to shoreline, and flood depth). Once the data was collected, average values for elevation and runoff parameters were calculated, which were then saved to the flood database. In addition, distance to coastline measurements were taken and saved in raster form for further analysis.

Next, the features used for training and testing the NNmodel were determined. These features include all the parameters that have been collected (elevation, surface flow, distance to coastline, and tide). The target is then the flood depth class. The data is then divided into training and testing sets to train the model and evaluate its performance.

The first layer has 64 neurons, and the second layer also has 32 neurons, both using the sigmoid activation function, which helps capture non-linear patterns in the data. The last layer contains one neuron with a linear activation function, which is used to produce continuous output. Backpropagation was performed using the ADAM optimizer the MSE loss function, and the model was trained for 50 epochs.

This step presents a flood hazard analysis for three regions: Jepara, Semarang, and Demak. The analysis consists of two primary components: an evaluation of the predictive model's performance, assessed using key classification metrics (Accuracy, Precision, Recall, and F1-Score) to ensure its reliability; and the calculation of Relative Feature Importance Weights, which ultimately serve as the final flood hazard weights for each location.

Table 5: Flood hazard weight

		Jepara	Semarang	Demak
Evaluation	Accuracy	0.9474	0.9333	0.8788
	Precision	0.9649	0.8800	0.8970
	Recall	0.9474	0.9333	0.8788
	F1-Score	0.9502	0.9037	0.8697
Relative Feature Importance Weight	Elevation	0.2278	0.2321	0.1679
	Runoff	0.3616	0.1998	0.3235
	Distance	0.2264	0.2775	0.3220
	Sea Level	0.1842	0.2906	0.1866

Based on the evaluation metrics (Table 5), the predictive models demonstrate results ranging from good to excellent across all three regions. The model for Jepara delivers the strongest outcome, achieving an exceptional F1-Score of 95.02% and an accuracy of 94.74%. This indicates an outstanding balance between precision and recall. Similarly, Semarang's model is also highly effective, yielding a strong F1-score of 90.37% and an accuracy of 93.33%. Meanwhile, the Demak model operates at a competent level, although slightly lower than the others, registering an F1-score of 86.97% and an accuracy of 87.88%. The high accuracy scores signify that the models are reliable in distinguishing between flood-prone and non-flood-prone areas. For instance, the 94.74% accuracy in Jepara means the model correctly classified approximately 95 out of every 100 conditions. It is important to note that while a predictive model was developed in this study, its primary output is the calculation of these feature weights, which will be utilized for hazard visualization and mapping.

Regarding the Determining Factors of Flood Hazard (Relative Feature Importance Weight), the analysis reveals that the primary drivers of flood risk vary by location, highlighting the unique characteristics of each region. In Jepara, flood hazard is primarily driven by Runoff (surface runoff), which holds a weight of 0.3616. For the Semarang region, the most significant variable is Sea Level (weight: 0.2906), closely followed by Distance from a river/sea at 0.2775. In Demak, the risk is most influenced by two nearly equally powerful factors: Runoff (0.3235) and Distance (0.3220). Overall, these results indicate that although all features (Elevation, Runoff, Distance, and Sea Level) are relevant, the mitigation priorities should differ for each regency. Jepara needs to focus on surface runoff management, while Semarang must prioritize mitigation related to sea-level rise, and Demak should address runoff and distance from water sources in a balanced manner.

The differences in weights reflect each location's unique environmental conditions and flood dynamics, which necessitate tailored approaches to flood risk analysis and management.

4.2 TES & NN combination to forecast panel data

To create the first model: forecasting with TES, starting with the processing of sea level and weather data. This data processing includes: converting dates from UTC to UTC+7, pre-processing each variable (converting sea level values from meters to centimeters, converting temperature from Kelvin to Celsius, etc.), grouping data by month for maximum values (sea level, V10, U10), average values (D2M, T2M), and minimum values (MSL, SP), then normalizing the data using the decimal scaling method. The forecasting results in three regions (Semarang, Demak, and Jepara) show different error rates for each variable. Table 6 presents the performance evaluation of the initial forecasting model, which was created as the first model in the study. The model's performance is measured using error metrics: MSE, RMSE, & MAPE, with lower error values indicating higher model accuracy. The results show the model's high accuracy (MAPE in 1.6%-2.2%) in predicting the weather variables: U10, V10, D2M, T2M, MSL, and SP. Conversely, the model faces significant challenges in predicting TP (MAPE≈70%), which has the highest error values, so this variable is excluded in the next process, also with SWH which actually has a fairly good error (MAPE≈18%) but not as high as other variables. For the Z variable (sea level) at three locations (Jepara, Semarang, and Demak), the prediction errors range from 22% to 25%, indicating moderate accuracy. Overall, this table indicates that the initial model is already very reliable for some meteorological variables but requires substantial improvements for other variables, especially rainfall and wave height. Accurate initial predictions for these variables are the reason they are included in the next stage. Z (sea level) has a moderate error rate, indicating that this model has fairly good accuracy in predicting sea level (Table 6, Figure 5), but it needs to consider the influence of weather variables to improve accuracy results.

Table 6: Initial Forecast, to create first model

Variables	MSE	RMSE	MAPE (%)
U10	0.0011	0.0327	1.67
V10	0.0008	0.0282	2.01
D2M	0.3895	0.6241	1.94
T2M	0.4683	0.6843	1.71
MSL	610.8699	24.7158	2.20
SP	626.0488	25.0210	2.23
TP	0.0062	0.0790	72.36
SWH	0.0253	0.1590	18.67
Z Jepara	174.7511	13.2193	22.87
Z Semarang	177.2125	13.3121	23.71
Z Demak	196.6765	14.0241	24.75

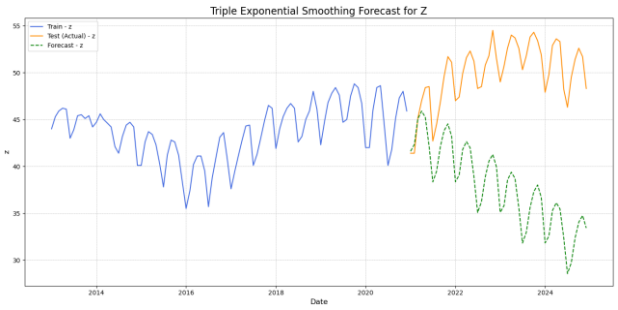


Figure 5: Initial forecast using TES

Training the NN model (for the second model, hybrid), by combining two methods: Triple Exponential Smoothing (TES) and Neural Network (NN), using sea level data influenced by several weather variables, starting with reading the data. Then, the data is divided into training data (1995-2020) and testing data (2021-2024). Next, two models are trained separately using the training data: (1) The TES model is trained on the time series data z to capture trend and seasonality patterns, (2) The NN model is trained to learn the relationship between selected feature variables (U10, V10, D2M, T2M, MSL, and SP) and the target variable z. These feature data are first scaled using MinMaxScaler. The NN model was created using MLPRegressor with six hidden layers containing 50 neurons.

The model performance evaluation (Table 7 & Figure 6) clearly shows that the Hybrid Method consistently outperforms the single method in all three regions. This superiority is evidenced by a drastic reduction in error values, with the MAPE error rate successfully reduced to less than half in all locations (e.g., from 20% to 4%). The primary reason behind this improvement in accuracy is that the hybrid method has higher accountability by incorporating the influence of weather variables into the modeling, a factor not considered by single methods. As a result, the hybrid approach produces forecasts that are not only more statistically accurate but also more reliable and comprehensive.

Table 7: Comparison between single method (TES) and hybrid method (TES & NN)

		Jepara	Semarang	Demak
Single Method	MSE	174.7511	177.2125	196.6765
	RMSE	13.2193	13.3121	14.0241
	MAPE	22.8745 %	23.7123 %	24.7514 %
Hybrid	MSE	25.1946	22.3257	22.6336
	RMSE	5.0194	4.7250	4.7575
	MAPE	8.6873%	8.3628%	8.3404%
Hybrid with Predicted Bias	MSE	4.8567	5.3047	5.0739
	RMSE	2.2038	2.3032	2.2525
	MAPE	3.5509%	3.7958%	3.6839%

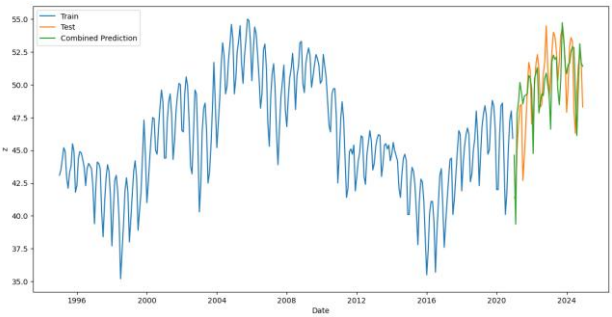


Figure 6: Forecast training & testing using hybrid method

4.3 Forecast with TES, NN & predicted bias

NN training was carried out to correct bias in time series data. The data was divided into training sets and test sets, with the features (X) and target (Y) used. The feature is then normalized using the MinMaxScaler into a range of 0 to 1. The NN model is made with three layers: an input layer with 64 neurons, a hidden layer with 32 neurons, and an output layer with 1 neuron. The model was compiled with the ADAM optimizer and MSE loss function, and then trained using training data that had been scaled over 50 epochs with a batch size of 32 and validation of 20% of the training data. The trained model is used to predict bias on the test set. The predicted average bias is calculated and adjusted by dividing by the number of features and factors.

Next, a combination of the TES and NN methods to make a combined prediction. The data was divided into training sets and test sets. The TES model is trained using training data and is used to make predictions on the test set. The NN model is trained with training data and is also used to make predictions on the test set. The predictions from both models were combined by taking

the mean from the predictions of the TES and the NN and then subtracting the bias. Actual data is combined with forecast data. MSE and MAPE are calculated to evaluate the performance of the model on the test set. In addition, a forecast for the 2024-2044 period was also made (Figure 7). In the context of this long-term flood hazard forecast, model deviations are of significant concern. To take this into account, this hybrid forecasting method allows for the integration of time series data and complex interactions between various influencing factors. This approach helps in adapting the model to better reflect changing conditions over time, thereby reducing the risk of model deviation.

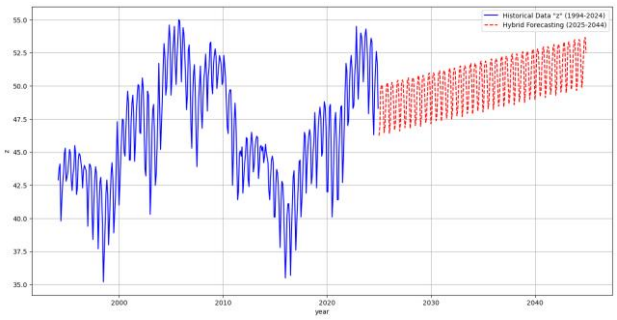


Figure 7: Sea Level Forecast 2025-2044

4.4 Visualization (Flood Hazard)

The visualization is provided in the form of a raster map created by calculating the flood hazard levels. Starting with creating a raster for elevation, several GeoTIFF files from DEMNAS, which form each region (Jepara/Semarang/Demak), are merged and then clipped according to the regional boundaries (shapefile of Indonesia's topographical boundary). The coordinate reference system (CRS) was changed to EPSG:4326, and the spatial resolution was converted to 8.35 meters. Then, for visualization purposes on the flood hazard map in the form of scores, the elevation was transformed using min-max normalization, and an inversion was performed (the higher the elevation, the lower the score).

Determination of CRS, clipping, and resampling to achieve the exact resolution is also performed on the landcover raster. The landcover values are converted to runoff based on the Rational Method Runoff Coefficient. This runoff value is considered as the score.

To create a raster distance to the coastline, the shapefile of the Indonesian Landform Boundary was used, and then the Euclidean distance transformation was calculated to determine the distance from each point of each pixel in the raster to the nearest coastline (which has been defined previously). This distance is then converted to meters. For score determination, min-max normalization is performed, and since this distance has a negative correlation with flood risk, inversion is carried out.

To create the sea level raster, each pixel is calculated: Sea level = Sea level Score × Distance Score, which means using the distance raster and linking it with the sea level value from the forecast (or actual value) at a specific time (month and year) determined as a parameter.

Finally, the flood hazard map is created by calculating the pixel values on all rasters associated with the predetermined weight values (different in each region).

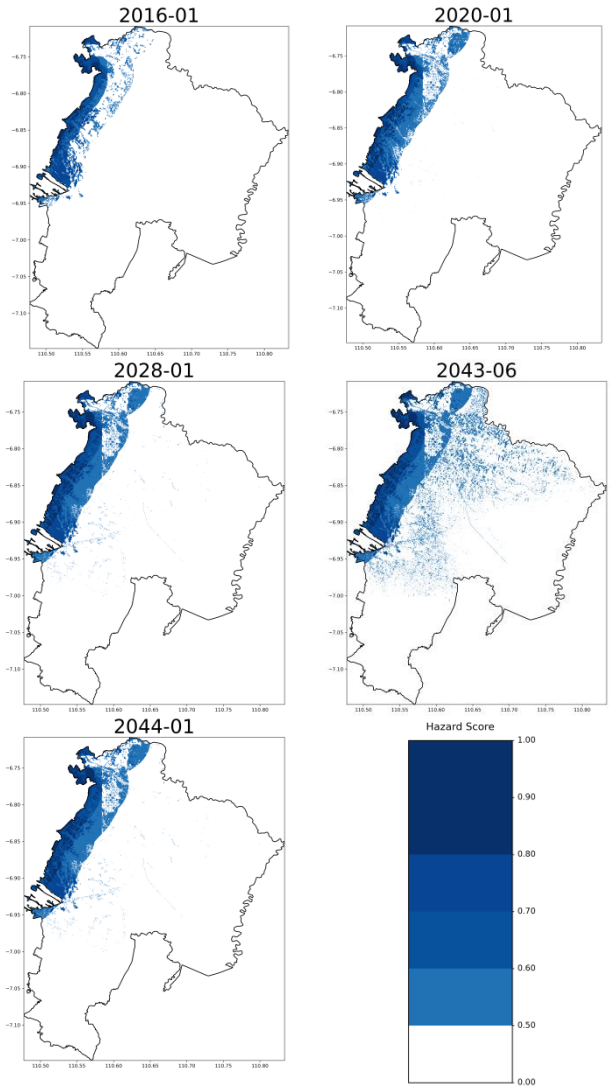


Figure 8: Flood Hazard: Demak

The projected future growth in flood-affected areas and increased hazard in the Demak area are depicted in Figure 8. Greater danger is indicated by darker blue hues, and the extent of the at-risk areas is depicted by larger blue regions. The northwestern coastal region was the primary location for flood hazards in January 2016 and January 2020. From 2016 to 2020, the number of flood-prone areas increased slightly, with areas at risk of flooding (light blue) seemingly moving a little farther inland. This means that the risk has already shown an upward trend during those four years. Predictions for January 2028 indicate a marked expansion of the at-hazard area in comparison to 2020. The light-blue patches extend farther into the interior, suggesting that areas that were previously low-hazard or not at risk are now starting to face the threat of flooding. In coastal

areas (dark blue), the level of hazard intensity is still high and has somewhat increased. The most dire scenario is projected for June 2043 (Long-Term Prediction), when areas at risk of flooding will have greatly expanded. Flood risk is expected to be low to high in nearly half of the northern portion of the region. According to this, the impact of climate change or other factors (like land subsidence and changes in land use) is expected to significantly worsen the flood situation within a 20-year period starting in 2028.

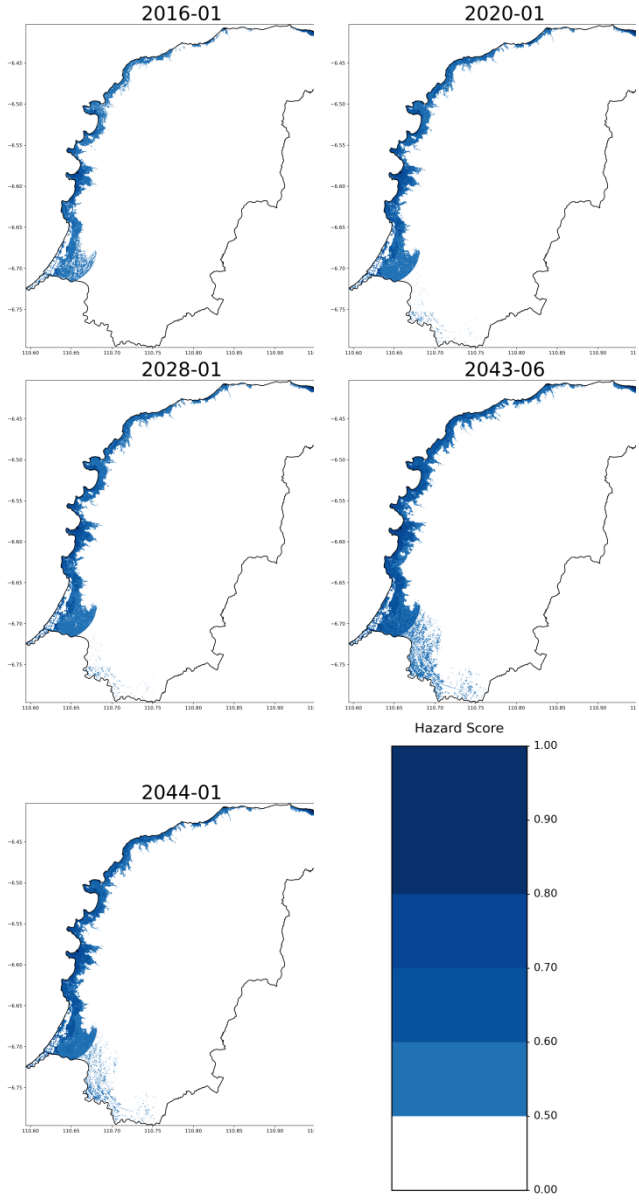


Figure 9: Flood Hazard: Jepara

The Jepara area, which has a long coastline, is predicted to have tidal flood or coastal inundation hazards, as shown in Figure 9. This set particularly emphasizes the dangers along the coast, in contrast to the Demak area, which depicts the spread of floods inland. Flood hazards were concentrated in the bay areas, river estuaries, and lowlands along the coastline during both of the 2016–2020 timeframes. The nearly identical hazard pattern during this time suggests that the risk of coastal inundation is comparatively constant. The most dangerous areas (dark blue) are found in the coastal basins that are most at risk. Hazardous area extent is predicted to slightly increase in 2028. The inundation areas (light blue) have slightly expanded at a number of locations along the coast, though not significantly, suggesting that the sea level may be rising or that high tides may be occurring more frequently. This long-term forecast shows a more pronounced increase in danger than in prior years. Particularly in the southern portion and in a number of coastal areas in the north, the inundation areas (light blue) seem to be more widespread. This suggests that low-lying coastal areas will experience more widespread and long-lasting flooding by 2043 as a result of the anticipated rise in sea levels. Wider low-risk zones now encircle the highest-risk areas (dark blue), which stay in the same place.

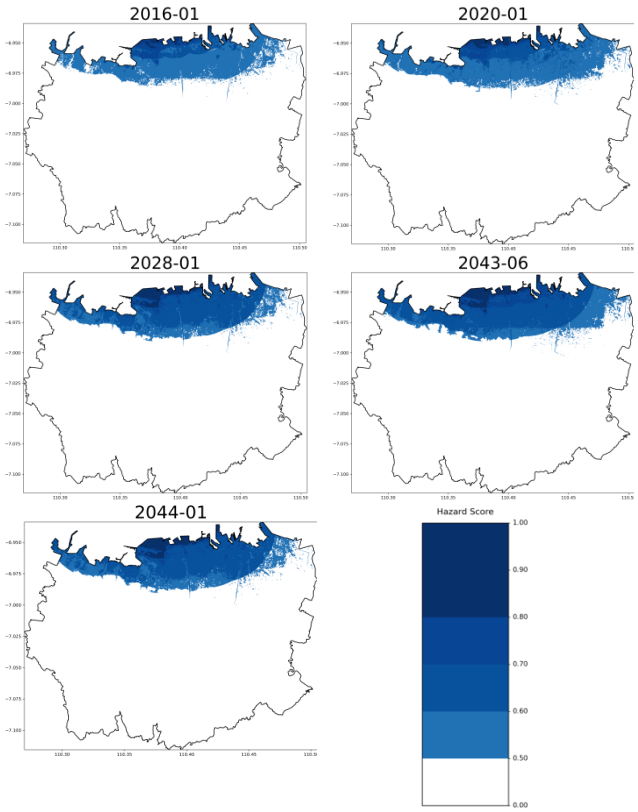


Figure 10: Flood Hazard: Semarang

The city of Semarang is seriously threatened by coastal flooding (rob), which is expected to get worse and spread more widely in the future, according to Figure 10. The majority of Semarang's northern coastal regions already had a high risk of flooding in 2016. While the light blue (lower risk) areas have already spread quite far southward (inland), the dark blue (highest risk) areas are concentrated close to the coast. Comparing 2020 to 2016, there was a minor decline. Flooded areas appear slightly larger and have spread farther inland, particularly those at lower risk (light blue). The impacted area is expected to significantly expand by 2028. A larger area to the

south is expected to be inundated by flooding than in 2020. Deeper or more frequent flooding is indicated by the dark blue areas' apparent slight increase in intensity. The Long-Term Prediction for 2043 is a worrying situation. The map displays a significant and striking increase in tidal flooding. It is anticipated that the inundation area will "consume" a significant portion of North Semarang and extend deep into the city center. Nearly the whole coastal region has turned into a high-risk zone (dark blue), meaning that flooding is likely to occur there either permanently or very frequently.

5 Discussion section

This study examines how rising sea levels made worse by climate change are increasing the risk of coastal floods in Indonesia's coastal regions, especially in Semarang, Demak, and Jepara, using spatial data mining forecasting techniques. Compared to conventional approaches, the hybrid model combining TES and NN significantly improves flood prediction accuracy. The hybrid model achieved a MAPE value of 8.3% to 3.5%, indicating that it is better at predicting floods more accurately compared to single methods [12]. Incorporating sea level data, elevation, runoff, distance to the coastline, and other parameters into the model allows for a comprehensive analysis of flood hazards. By using historical flood data to measure the weight of these parameters through NN, the assessment becomes more objective and subjectivity in the Analytic Hierarchy Process (AHP) is reduced [8]. For areas like Indonesia, where geographical and climatic variances can greatly affect flood dynamics, this methodological development is essential. This study also highlights how crucial it is to analyze very large datasets using spatial data mining to spot trends and raise the precision of predictive models. In addition to identifying non-linear correlations, the hybrid approach is expected to overcome the weaknesses of traditional forecasting techniques, which often ignore the influences between variables [6]. The results indicate that the risk of flooding will increase in hundreds of densely populated cities by 2050, potentially doubling by 2100 [1].

The trade-off between NN, exponential smoothing, and hybrid approaches (Table 10) involves many considerations. NN provide scalability and the ability to model complex relationships at the expense of computational efficiency and interpretability. Exponential Smoothing offers a more straightforward and interpretable approach that is computationally efficient but may struggle with scalability and complexity. Hybrid Approaches aim to combine the strengths of both methods, providing a balanced solution that improves scalability, maintains a degree of interpretability, and improves accuracy.

Table 10: Trade-offs between all the methods used

Aspect	NN	ES	Hybrid
Computational Efficiency	High computational demand; longer training times	Low computational demand; quick execution	High computational demand; resource-intensive
Scalability	Highly scalable; adapts to large datasets	Limited scalability; struggles with complexity	Highly scalable; flexible adaptation
Interpretability	Low interpretability (black box nature)	High interpretability; straightforward mechanics	Moderate interpretability; combines strengths
Model Accuracy	High accuracy potential; risk of overfitting	Good accuracy for simple patterns; limited in complexity	Enhanced accuracy; robust to variability

Based on the weight analysis derived from Feature importance (Table 5), it can be concluded that the main triggers of coastal flooding vary significantly in each location, indicating unique local vulnerabilities [116]. High water runoff from the land poses the biggest risk in Jepara. Semarang, on the other hand, is more susceptible to direct sea influences, especially those related to sea level and land subsidence (elevation). In the meantime, water runoff and Demak's closeness to the coast pose a nearly equal threat. This key distinction highlights the fact that mitigation strategies must be tailored to each region and cannot be applied universally.

6 Conclusion

The hybrid method applied has successfully addressed the panel data problem with relatively better MAPE values, ranging from 8.3% to 3.5%. Additionally, the flood hazard weights have also been well identified, with accuracy values varying between 88% and 94%, regional differences can affect results due to the different geographical and climatic characteristics in each region, which can affect flood dynamics.

The findings of this study contribute to the increasing knowledge of flood risk assessment and management, providing a solid framework for future studies and practical applications in flood-prone areas.

The success of the hybrid model in predicting coastal floods can serve as a valuable tool for policymakers and urban planners in developing effective flood management strategies. It is hoped that decision-makers can benefit from appropriate flood mitigation techniques for coastal areas, as infrastructure and lives are seriously threatened by this condition.

Some points that can be concluded:

1. Integrating a hybrid approach with climate model projections is recommended in climate change uncertainty with significant variations due to differences in physics assumptions, parameterization, and emission scenarios by capturing a more comprehensive range of possible future conditions.
2. Predictions become more resilient to biases or model-specific errors by using multiple projections at once, resulting in more reliable estimates of flood risk or related variables under climate change scenarios.
3. Probabilistic information from hybrids supporting risk-based decision-making can be used to design flexible adaptation strategies for future possibilities.

This research opens up essential insights and broader implications for coastal policy and risk management:

1. Scalability of the System
   Hybrid models can be widely applied in various coastal areas with different characteristics as long as adequate data is available.
2. Real-World Application
   The model must be followed by real implementation, including integration with early warning systems and regional spatial planning.
3. Implementation Challenges
   Field implementation may face data limitations, human resources, and technological infrastructure. Therefore, adequate policy support and investment must be needed to overcome these barriers.

7 Acknowledgements

This study received support from Doctoral Program of Information System, School of Postgraduate Studies, Universitas Diponegoro.

References

[1] UNDP, "Climate change's impact on coastal flooding to increase 5-times over this century, putting over 70 million people in the path of expanding floodplains, according to new UNDP and CIL data." Accessed: Nov. 25, 2024. [Online]. Available: https://www.undp.org/press-releases/climate-changes-impact-coastal-flooding-increase-5-times-over-century-putting-over-70-million-people-path-expanding-floodplains

[2] V. Kumar, K. Sharma, T. Caloiero, D. Mehta, and K. Singh, "Comprehensive Overview of Flood Modeling Approaches: A Review of Recent Advances," *Hydrology*, vol. 10, no. 7, p. 141, Jun. 2023, doi: 10.3390/hydrology10070141.

[3] F. Aureli, A. Maranzoni, and G. Petaccia, "Advances in Dam-Break Modeling for Flood Hazard Mitigation: Theory, Numerical Models, and Applications in Hydraulic Engineering," *Water*, vol. 16, no. 8, p. 1093, Apr. 2024, doi: 10.3390/w16081093.

[4] P. Sibandze, A. M. Kalumba, A. H. Aljaddani, L. Zhou, and G. A. Afuye, "Geospatial Mapping and Meteorological Flood Risk Assessment: A Global Research Trend Analysis," *Environ. Manage.*, Oct. 2024, doi: 10.1007/s00267-024-02059-0.

[5] FEMA, "Guidance for Flood Risk Analysis and Mapping." [Online]. Available: https://www.fema.gov/sites/default/files/documents/fema_flood-risk-assessment-guidance.pdf

[6] J. Lin, C. Sreng, E. Oare, and F. A. Batarseh, "NeuralFlood: an AI-driven flood susceptibility index," *Front. Water*, vol. 5, Oct. 2023, doi: 10.3389/frwa.2023.1291305.

[7] M. Lang *et al.*, "Use of Systematic, Palaeoflood and Historical Data for the Improvement of Flood Risk Estimation. Review of Scientific Methods," *Nat. Hazards*, vol. 31, no. 3, pp. 623–643, Mar. 2004, doi: 10.1023/B:NHAZ.0000024895.48463.eb.

[8] K. Yao *et al.*, "A Novel Flood Risk Analysis Framework Based on Earth Observation Data to Retrieve Historical Inundations and Future Scenarios," *Remote Sens.*, vol. 16, no. 8, p. 1413, Apr. 2024, doi: 10.3390/rs16081413.

[9] Y. Chen, D. Wang, L. Zhang, H. Guo, J. Ma, and W. Gao, "Flood risk assessment of Wuhan, China, using a multi-criteria analysis model with the improved AHP-Entropy method," *Environ. Sci. Pollut. Res.*, vol. 30, no. 42, pp. 96001–96018, Aug. 2023, doi: 10.1007/s11356-023-29066-8.

[10] S. R. Singh, E. Harirchian, C. E. F. Monjardin, and T. Lahmer, "GIS-Based Risk Assessment of Building Vulnerability in Flood Zones of Naic, Cavite, Philippines Using AHP and TOPSIS," *GeoHazards*, vol. 5, no. 4, pp. 1040–1073, Oct. 2024, doi: 10.3390/geohazards5040050.

[11] R. Oyelakin, W. Yang, and P. Krebs, "Analysing Urban Flooding Risk with CMIP5 and CMIP6 Climate Projections," *Water*, vol. 16, no. 3, p. 474, Jan. 2024, doi: 10.3390/w16030474.

[12] Z. Yin *et al.*, "Assessing the economic impacts of future fluvial flooding in six countries under climate change and socio-economic development," *Clim. Change*, vol. 166, no. 3–4, 2021, doi: 10.1007/s10584-021-03059-3.

[13] D. K. Hakim, R. Gernowo, and A. W.

Nirwansyah, "Flood prediction with time series data mining: Systematic review," *Nat. Hazards Res.*, vol. 4, no. 2, pp. 194–220, Oct. 2023, doi: 10.1016/j.nhres.2023.10.001.

[14] R. Lan, "Spatial Data Mining and Knowledge Discovery," in *Advances in Cartography and Geographic Information Engineering*, Singapore: Springer Singapore, 2021, pp. 503–530. doi: 10.1007/978-981-16-0614-4_14.

[15] J. Henson, "Univariate vs Multivariate Time Series Forecasting." Accessed: Oct. 18, 2023. [Online]. Available: https://medium.com/@jesse.henson/univariate-vs-multivariate-time-series-forecasting-cfcc4150e20a

[16] J. Brownlee, "Taxonomy of Time Series Forecasting Problems," Machine Learning Mastery. Accessed: Oct. 18, 2023. [Online]. Available: https://machinelearningmastery.com/taxonomy-of-time-series-forecasting-problems/

[17] F. Petropoulos and E. Spiliotis, "The Wisdom of the Data: Getting the Most Out of Univariate Time Series Forecasting," *Forecasting*, vol. 3, no. 3, pp. 478–497, Jun. 2021, doi: 10.3390/forecast3030029.

[18] N. Rankovic, D. Rankovic, M. Ivanovic, and L. Lazic, "Improved effort and cost estimation model using artificial neural networks and taguchi method with different activation functions," *Entropy*, vol. 23, no. 7, 2021, doi: 10.3390/e23070854.

[19] A. Sahoo, S. Samantaray, and D. K. K. Ghose, "Prediction of Flood in Barak River using Hybrid Machine Learning Approaches: A Case Study," *J. Geol. Soc. India*, vol. 97, no. 2, pp. 186–198, 2021, doi: 10.1007/s12594-021-1650-1.

[20] A. H. Zuhairi, F. Yakub, S. A. Zaki, and M. S. M. Ali, "Review of flood prediction hybrid machine learning models using datasets," 2022. doi: 10.1088/1755-1315/1091/1/012040.

[21] I. Ardiansah *et al.*, "Integrated Streamflow Forecasting System: A Step Towards Smart Flood Management," *Inform.*, vol. 47, no. 9, pp. 109–121, 2023, doi: 10.31449/inf.v47i9.4890.

[22] T. Canchala, W. Alfonso-Morales, Y. Carvajal-Escobar, W. L. Cerón, and E. Caicedo-Bravo, "Monthly Rainfall Anomalies Forecasting for Southwestern Colombia Using Artificial Neural Networks Approaches," *Water*, vol. 12, no. 9, p. 2628, Sep. 2020, doi: 10.3390/w12092628.

[23] K. Xu, Z. Han, L. Bin, R. Shen, and Y. Long, "Rapid forecasting of compound flooding for a coastal area based on data-driven approach," *Nat. Hazards*, vol. 121, no. 2, pp. 1399–1421, Jan. 2025, doi: 10.1007/s11069-024-06846-0.

[24] B. Victor *et al.*, "Off to new Shores: A Dataset & Benchmark for (near-)coastal Flood Inundation Forecasting," 2024. [Online]. Available: https://arxiv.org/abs/2409.18591

[25] J. Chou, Y. Cao, and P. Minford, "Evaluation and indirect inference estimation of inattentive features in a New Keynesian framework," *J. Forecast.*, vol. 42, no. 3, pp. 530–542, Apr. 2023, doi: 10.1002/for.2950.

[26] X.-P. Li *et al.*, "The Study of Biological Glue Droplet Impact Behavior of Bioceramic Powders Applied in 3D Printing of Bone Scaffolds," *Appl. Sci.*, vol. 12, no. 4, p. 1898, Feb. 2022, doi: 10.3390/app12041898.

[27] M. Habulan, B. Đurin, A. P. Siročić, and N. Sakač, "Analysis of the Particulate Matter Pollution in the Urban Areas of Croatia, EU," in *The 3rd International Electronic Conference on Atmospheric Sciences*, Basel Switzerland: MDPI, Nov. 2020, p. 9. doi: 10.3390/ecas2020-08145.

[28] M.-J. Park and H.-S. Yang, "Comparative Study of Time Series Analysis Algorithms Suitable for Short-Term Forecasting in Implementing Demand Response Based on AMI," *Sensors*, vol. 24, no. 22, p. 7205, Nov. 2024, doi: 10.3390/s24227205.

[29] R. V. Vargas and P. M. P. IPMA-B, "Using the analytic hierarchy process (AHP) to select and prioritize projects in a portfolio," in *PMI global congress*, PA: Project Management Institute Washington, DC, 2010, pp. 1–22.

[30] M. Tavana, M. Soltanifar, and F. J. Santos-Arteaga, "Analytical hierarchy process: revolution and evolution," *Ann. Oper. Res.*, vol. 326, no. 2, pp. 879–907, Jul. 2023, doi: 10.1007/s10479-021-04432-2.

[31] P. Madzík and L. Falát, "State-of-the-art on analytic hierarchy process in the last 40 years: Literature review based on Latent Dirichlet Allocation topic modelling," *PLoS One*, vol. 17, no. 5, p. e0268777, May 2022, doi: 10.1371/journal.pone.0268777.

[32] S. Mourato, P. Fernandez, L. G. Pereira, and M. Moreira, "Assessing Vulnerability in Flood Prone Areas Using Analytic Hierarchy Process—Group Decision Making and Geographic Information System: A Case Study in Portugal," *Appl. Sci.*, vol. 13, no. 8, p. 4915, Apr. 2023, doi: 10.3390/app13084915.

[33] K. C. Swain, C. Singha, and L. Nayak, "Flood Susceptibility Mapping through the GIS-AHP Technique Using the Cloud," *ISPRS Int. J. Geo-Information*, vol. 9, no. 12, p. 720, Dec. 2020, doi: 10.3390/ijgi9120720.

[34] BPS, "Provider of Quality Statistical Data For Advanced Indonesia." Accessed: Dec. 05, 2024. [Online]. Available: https://www.bps.go.id/id

[35] E. Hermawan *et al.*, "Large-Scale Meteorological Drivers of the Extreme Precipitation Event and Devastating Floods of Early-February 2021 in Semarang, Central Java, Indonesia," *Atmosphere (Basel).*, vol. 13, no. 7, 2022, doi: 10.3390/atmos13071092.

[36] A. Chrysanti, A. Adhani, I. N. Azkiarizqi, M. B. Adityawan, M. S. B. Kusuma, and M. Cahyono, "Assessing Compound Coastal–Fluvial Flood

Impacts and Resilience Under Extreme Scenarios in Demak, Indonesia," *Sustainability*, vol. 16, no. 23, p. 10315, Nov. 2024, doi: 10.3390/su162310315.

[37] N. A. Akhirianto, K. S. Wardani, A. Ma'rufatin, A. Nurwijayanti, F. M. G. Putra, and F. Chusnayah, "Sustainable regional development planning based on coastal disaster risk assessment (Case study: North coast of Central Java)," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1479, no. 1, p. 012040, Apr. 2025, doi: 10.1088/1755-1315/1479/1/012040.

[38] M. Imani *et al.*, "Risk Assessment of Coastal Flooding under Different Inundation Situations in Southwest of Taiwan (Tainan City)," *Water*, vol. 13, no. 6, p. 880, Mar. 2021, doi: 10.3390/w13060880.

[39] M. Tu'uholoaki, A. Espejo, M. Wandres, A. Singh, H. Damlamian, and Z. Begg, "Quantifying Mechanisms Responsible for Extreme Coastal Water Levels and Flooding during Severe Tropical Cyclone Harold in Tonga, Southwest Pacific," *J. Mar. Sci. Eng.*, vol. 11, no. 6, p. 1217, Jun. 2023, doi: 10.3390/jmse11061217.

[40] J.-H. Yuk, J.-S. Kang, and H. Myung, "Applicability Study of a Global Numerical Weather Prediction Model MPAS to Storm Surges and Waves in the South Coast of Korea," *Atmosphere (Basel).*, vol. 13, no. 4, 2022, doi: 10.3390/atmos13040591.

[41] W. Costa, D. Idier, J. Rohmer, M. Menendez, and P. Camus, "Statistical Prediction of Extreme Storm Surges Based on a Fully Supervised Weather-Type Downscaling Model," *J. Mar. Sci. Eng.*, vol. 8, no. 12, p. 1028, Dec. 2020, doi: 10.3390/jmse8121028.

[42] L. Fan and L. Du, "Combined effects of climatic factors on extreme sea level changes in the Northwest Pacific Ocean," *Ocean Dyn.*, vol. 73, no. 3–4, pp. 181–199, 2023, doi: 10.1007/s10236-023-01543-1.

[43] M. Buchhorn, M. Lesiv, N.-E. Tsendbazar, M. Herold, L. Bertels, and B. Smets, "Copernicus Global Land Cover Layers—Collection 2," *Remote Sens.*, vol. 12, no. 6, p. 1044, Mar. 2020, doi: 10.3390/rs12061044.

[44] CivilWeb, "Rational Method Runoff Coefficient." Accessed: Aug. 09, 2024. [Online]. Available: https://civilweb-spreadsheets.com/drainage-design-spreadsheets/runoff-and-rainfall-intensity-calculator-spreadsheet/rational-method-runoff-coefficient/

[45] SRGI, "Tide Prediction: Indonesian Geospatial Reference System." Accessed: Nov. 01, 2024. [Online]. Available: https://srgi.big.go.id/map/pasut-prediction-active

[46] J.-N. Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, "ERA5 hourly data on single levels from 1940 to present," 2023. doi: 10.24381/cds.adbb2d47.

[47] D. F. Hill, "Spatial and Temporal Variability in Tidal Range: Evidence, Causes, and Effects," *Curr. Clim. Chang. Reports*, vol. 2, no. 4, pp. 232–241, Dec. 2016, doi: 10.1007/s40641-016-0044-8.

[48] A. T. Devlin, D. A. Jay, E. D. Zaron, S. A. Talke, J. Pan, and H. Lin, "Tidal Variability Related to Sea Level Variability in the Pacific Ocean," *J. Geophys. Res. Ocean.*, vol. 122, no. 11, pp. 8445–8463, Nov. 2017, doi: 10.1002/2017JC013165.

[49] J. Challis, D. Idier, G. Wöppelmann, and G. André, "Atmospheric Wind and Pressure-Driven Changes in Tidal Characteristics over the Northwestern European Shelf," *J. Mar. Sci. Eng.*, vol. 11, no. 9, p. 1701, Aug. 2023, doi: 10.3390/jmse11091701.

[50] Copernicus Climate Change Service, "ERA5: data documentation," 2025. [Online]. Available: https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation

[51] Copernicus Climate Change Service, "ERA5: What is the spatial reference," 2025. [Online]. Available: https://confluence.ecmwf.int/display/CKB/ERA5%3A+What+is+the+spatial+reference

[52] K. Park and E. H. Lee, "Urban flood vulnerability analysis and prediction based on the land use using Deep Neural Network," *Int. J. Disaster Risk Reduct.*, vol. 101, p. 104231, Feb. 2024, doi: 10.1016/j.ijdrr.2023.104231.

[53] A. Nirwansyah, *Geodata dan Pemanfaatannya dalam Isu Kebencanaan*. 2024.

[54] M. Lorie *et al.*, "Modeling coastal flood risk and adaptation response under future climate conditions," *Clim. Risk Manag.*, vol. 29, p. 100233, 2020, doi: https://doi.org/10.1016/j.crm.2020.100233.

[55] H. Moftakhari, J. E. Schubert, A. AghaKouchak, R. A. Matthew, and B. F. Sanders, "Linking statistical and hydrodynamic modeling for compound flood hazard assessment in tidal channels and estuaries," *Adv. Water Resour.*, vol. 128, pp. 28–38, Jun. 2019, doi: 10.1016/j.advwatres.2019.04.009.

[56] J. A. Pollard, T. Spencer, and S. Jude, "Big Data Approaches for coastal flood risk assessment and emergency response," *WIREs Clim. Chang.*, vol. 9, no. 5, pp. 1–14, Sep. 2018, doi: 10.1002/wcc.543.

[57] M. U. Parodi, A. Giardino, A. van Dongeren, S. G. Pearson, J. D. Bricker, and A. J. H. M. Reniers, "Uncertainties in coastal flood risk assessments in small island developing states," *Nat. Hazards Earth Syst. Sci.*, vol. 20, no. 9, pp. 2397–2414, Sep. 2020, doi: 10.5194/nhess-20-2397-2020.

[58] D. L. Anderson *et al.*, "Projecting Climate Dependent Coastal Flood Risk With a Hybrid Statistical Dynamical Model," *Earth's Futur.*, vol. 9, no. 12, 2021, doi: 10.1029/2021EF002285.

[59] A.-L. Balogun and N. Adebisi, "Sea level prediction using ARIMA, SVR and LSTM neural network: assessing the impact of ensemble Ocean-Atmospheric processes on models' accuracy," *Geomatics, Nat. Hazards Risk*, vol. 12, no. 1, pp. 653–674, Jan. 2021, doi: 10.1080/19475705.2021.1887372.

[60] Y. N. Hilal, G. D. A. Nainggolan, S. H. Syahputri, and F. Kartiasih, "Comparison of ARIM and LSTM Methods in Predicting Jakarta Sea Level," *J. Ilmu dan Teknol. Kelaut. Trop.*, vol. 16, no. 2, pp. 163–178, Oct. 2024, doi: 10.29244/jitkt.v16i2.52818.

[61] M. F. Marsani, V. Someetheram, M. S. Mohd Kasihmuddin, S. Z. Mohd Jamaludin, M. A. Mansor, and B. Badyalina, "The application of seasonal autoregressive integrated moving average (SARIMA) model in forecasting Malaysia mean sea level," 2024, p. 070005. doi: 10.1063/5.0223836.

[62] E. Ayitey, F. Ayiah-Mensah, S. Nunoo, and J. A. Addor, "Applying machine learning techniques for sea level rise forecasting in Axim: tackling missing data and outliers," *Model. Earth Syst. Environ.*, vol. 11, no. 3, p. 164, Jun. 2025, doi: 10.1007/s40808-025-02314-1.

[63] X. Long *et al.*, "Evaluating Current Statistical and Dynamical Forecasting Techniques for Seasonal Coastal Sea Level Prediction," *J. Clim.*, vol. 38, no. 6, pp. 1477–1503, Mar. 2025, doi: 10.1175/JCLI-D-24-0214.1.

[64] L. Elneel, M. S. Zitouni, H. Mukhtar, and H. Al-Ahmad, "Examining sea levels forecasting using autoregressive and prophet models," *Sci. Rep.*, vol. 14, no. 1, p. 14337, Jun. 2024, doi: 10.1038/s41598-024-65184-0.

[65] P. K. Srivastava, T. Islam, S. K. Singh, G. P. Petropoulos, M. Gupta, and Q. Dai, "Forecasting Arabian Sea level rise using exponential smoothing state space models and ARIMA from TOPEX and Jason satellite radar altimeter data," *Meteorol. Appl.*, vol. 23, no. 4, pp. 633–639, Oct. 2016, doi: 10.1002/met.1585.

[66] M. Jahanbani, M. H. Vahidnia, H. Aghamohammadi, and Z. Azizi, "Flood susceptibility mapping through geoinformatics and ensemble learning methods, with an emphasis on the AdaBoost-Decision Tree algorithm, in Mazandaran, Iran," *Earth Sci. Informatics*, vol. 17, no. 2, pp. 1433–1457, Apr. 2024, doi: 10.1007/s12145-023-01213-2.

[67] G. D. Kumar, S. Tyagi, and K. C. Pradhan, "Predictive ML analysis: Rainfall & flood vulnerability in Bihar, India," in *Artificial Intelligence and Information Technologies*, London: CRC Press, 2024, pp. 447–453. doi: 10.1201/9781032700502-71.

[68] G. D. Kumar, K. C. Pradhan, and S. Tyagi, "Deep Learning Forecasting: An LSTM Neural Architecture based Approach to Rainfall and Flood Impact Predictions in Bihar," *Procedia Comput. Sci.*, vol. 235, pp. 1455–1466, 2024, doi: 10.1016/j.procs.2024.04.137.

[69] D. K. Hakim, R. Gernowo, A. W. Nirwansyah, and T. Haryanto, "Data Mining Approach for River Flood Hazard Time-Series: Using a Combination of Triple Exponential Smoothing and Neural Networks, in Demak," *Informatica*, vol. 49, no. 19, Apr. 2025, doi: 10.31449/inf.v49i19.7288.

[70] BNPB, *IRBI 2023*. 2023. [Online]. Available: https://inarisk.bnpb.go.id/IRBI-2023/mobile/index.html

[71] C. Buta, G. Mihai, and M. Stănescu, "Flood Risk Assessment Based on Flood Hazard and Vulnerability Indexes," *Ovidius Univ. Ann. Constanta - Ser. Civ. Eng.*, vol. 22, no. 1, pp. 127–137, Dec. 2020, doi: 10.2478/ouacsce-2020-0014.

[72] FEMA, "Guidance for Flood Risk Analysis and Mapping. Hydrology: Rainfall-Runoff Analysis," 2019. [Online]. Available: https://www.fema.gov/sites/default/files/2020-02/Hydrologic_Rainfall_Runoff_Analysis_Feb_2019.pdf

[73] T. Tingsanchali and T. Promping, "Comprehensive Assessment of Flood Hazard, Vulnerability, and Flood Risk at the Household Level in a Municipality Area: A Case Study of Nan Province, Thailand," *Water*, vol. 14, no. 2, p. 161, Jan. 2022, doi: 10.3390/w14020161.

[74] UNISDR, "Flood Hazard and Risk Assessment," 2017. [Online]. Available: https://www.unisdr.org/files/52828_04floodhazardandriskassessment.pdf

[75] Q. Wang, H. Gu, X. Zang, M. Zuo, and H. Li, "Flood resilience in cities and urban agglomerations: a systematic review of hazard causes, assessment frameworks, and recovery strategies based on LLM tools," *Nat. Hazards*, May 2025, doi: 10.1007/s11069-025-07285-1.

[76] S. Zhang, Z. Yang, and C. Wang, "Environmental risk assessment of earthen heritage sites in arid regions: a case study of the Great Wall in Xinjiang," *npj Herit. Sci.*, vol. 13, no. 1, p. 200, May 2025, doi: 10.1038/s40494-025-01736-5.

[77] C. R. de Sá, "Variance-Based Feature Importance in Neural Networks," 2019, pp. 306–315. doi: 10.1007/978-3-030-33778-0_24.

[78] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Appl. Sci.*, vol. 3, no. 2, p. 272, Feb. 2021, doi: 10.1007/s42452-021-04148-9.

[79] K. Xu, J. Fang, Y. Fang, Q. Sun, C. Wu, and M. Liu, "The Importance of Digital Elevation Model

Selection in Flood Simulation and a Proposed Method to Reduce DEM Errors: A Case Study in Shanghai," *Int. J. Disaster Risk Sci.*, vol. 12, no. 6, pp. 890–902, 2021, doi: 10.1007/s13753-021-00377-z.

[80] G. Schumann, L. Giustarini, A. Tarpanelli, B. Jarihani, and S. Martinis, "Flood Modeling and Prediction Using Earth Observation Data," *Surv. Geophys.*, vol. 44, no. 5, pp. 1553–1578, Oct. 2023, doi: 10.1007/s10712-022-09751-y.

[81] D. B. Gesch, "Best Practices for Elevation-Based Assessments of Sea-Level Rise and Coastal Flooding Exposure," *Front. Earth Sci.*, vol. 6, Dec. 2018, doi: 10.3389/feart.2018.00230.

[82] H. Liang and X. Zhou, "Impact of Tides and Surges on Fluvial Floods in Coastal Regions," *Remote Sens.*, vol. 14, no. 22, 2022, doi: 10.3390/rs14225779.

[83] H. Zhou, *Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods*. Berkeley, CA: Apress, 2020. doi: 10.1007/978-1-4842-5982-5.

[84] R. Abdulkadirov, P. Lyakhov, and N. Nagornov, "Survey of Optimization Algorithms in Modern Neural Networks," *Mathematics*, vol. 11, no. 11, p. 2466, May 2023, doi: 10.3390/math11112466.

[85] S. Prasad, "Time Series," in *Advanced Statistical Methods*, Singapore: Springer Nature Singapore, 2024, pp. 109–162. doi: 10.1007/978-981-99-7257-9_3.

[86] C. Fieberg, D. Metko, T. Poddig, and T. Loy, "Machine learning techniques for cross-sectional equity returns' prediction," *OR Spectr.*, vol. 45, no. 1, pp. 289–323, Mar. 2023, doi: 10.1007/s00291-022-00693-w.

[87] F. Duval, J.-P. Boucher, and M. Pigeon, "Telematics combined actuarial neural networks for cross-sectional and longitudinal claim count data," *ASTIN Bull.*, vol. 54, no. 2, pp. 239–262, May 2024, doi: 10.1017/asb.2024.4.

[88] N. S. Grigg, "Comprehensive Flood Risk Assessment: State of the Practice," *Hydrology*, vol. 10, no. 2, p. 46, Feb. 2023, doi: 10.3390/hydrology10020046.

[89] A. Díez-Herrero and J. Garrote, "Flood Risk Assessments: Applications and Uncertainties," *Water*, vol. 12, no. 8, p. 2096, Jul. 2020, doi: 10.3390/w12082096.

[90] A. Buckle, "Why is it Called UTC – not CUT," timeanddate.com. Accessed: Jan. 02, 2025. [Online]. Available: https://www.timeanddate.com/time/utc-abbreviation.html

[91] A. Babii, R. T. Ball, E. Ghysels, and J. Striaukas, "Panel Data Nowcasting: The Case of Price-Earnings Ratios," 2023. [Online]. Available: https://arxiv.org/abs/2307.02673

[92] S. F. Crone and N. Kourentzes, "Feature selection for time series prediction – A combined filter and wrapper approach for neural networks," *Neurocomputing*, vol. 73, no. 10–12, pp. 1923–1936, Jun. 2010, doi: 10.1016/j.neucom.2010.01.017.

[93] N. J. Clark and K. Wells, "Dynamic generalised additive models (DGAMs) for forecasting discrete ecological time series," *Methods Ecol. Evol.*, vol. 14, no. 3, pp. 771–784, Mar. 2023, doi: 10.1111/2041-210X.13974.

[94] F. Kronthaler, *Statistics Applied With Excel*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023. doi: 10.1007/978-3-662-64319-8.

[95] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization Methods for Large-Scale Machine Learning," *SIAM Rev.*, vol. 60, no. 2, pp. 223–311, Jan. 2018, doi: 10.1137/16M1080173.

[96] M. T. Hagan, H. B. Demuth, and M. Beale, *Neural Network Design*, 2nd ed. Martin Hagan, 2014.

[97] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of Artificial Neural Networks and Deep Learning," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 379–425. doi: 10.1007/978-3-030-89010-0_10.

[98] A. A. Waoo and B. K. Soni, "Performance Analysis of Sigmoid and Relu Activation Functions in Deep Neural Network," 2021, pp. 39–52. doi: 10.1007/978-981-16-2248-9_5.

[99] J. Urenda, O. Kosheleva, and V. Kreinovich, "Fuzzy Techniques Explain the Effectiveness of ReLU Activation Function in Deep Learning," 2024, pp. 425–430. doi: 10.1007/978-3-031-55684-5_28.

[100] Y. Kim, I. Ohn, and D. Kim, "Fast convergence rates of deep neural networks for classification," 2019. [Online]. Available: https://arxiv.org/abs/1812.03599

[101] V. M. Vargas, D. Guijo-Rubio, P. A. Gutiérrez, and C. Hervás-Martínez, "ReLU-Based Activations: Analysis and Experimental Study for Deep Learning," 2021, pp. 33–43. doi: 10.1007/978-3-030-85713-4_4.

[102] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," 2020. [Online]. Available: https://arxiv.org/abs/1905.10437

[103] M. Shah and N. Sureja, "A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions," *Arch. Comput. Methods Eng.*, vol. 32, no. 1, pp. 255–267, Jan. 2025, doi: 10.1007/s11831-024-10134-2.

[104] H. Abbasimehr and R. Paki, "Improving time series forecasting using LSTM and attention models," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 1, pp. 673–691, Jan. 2022, doi: 10.1007/s12652-020-02761-x.

[105] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on Machine Learning Biases and Mitigation

Techniques," *Digital*, vol. 4, no. 1, pp. 1–68, Dec. 2023, doi: 10.3390/digital4010001.

[106] R. Wang, P. Chaudhari, and C. Davatzikos, "Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies," *Proc. Natl. Acad. Sci.*, vol. 120, no. 6, Feb. 2023, doi: 10.1073/pnas.2211613120.

[107] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice, 3rd edition*. Melbourne, Australia.: OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/

[108] M. M. Msabi and M. Makonyo, "Flood susceptibility mapping using GIS and multi-criteria decision analysis: A case of Dodoma region, central Tanzania," *Remote Sens. Appl. Soc. Environ.*, vol. 21, p. 100445, Jan. 2021, doi: 10.1016/j.rsase.2020.100445.

[109] D. Kvočka, R. A. Falconer, and M. Bray, "Flood hazard assessment for extreme flood events," *Nat. Hazards*, vol. 84, no. 3, pp. 1569–1599, Dec. 2016, doi: 10.1007/s11069-016-2501-z.

[110] N. Tabasi, M. Fereshtehpour, and B. Roghani, "A review of flood risk assessment frameworks and the development of hierarchical structures for risk components," *Discov. Water*, vol. 5, no. 1, p. 10, Feb. 2025, doi: 10.1007/s43832-025-00193-2.

[111] M. Pronk *et al.*, "DeltaDTM: A global coastal digital terrain model," *Sci. Data*, vol. 11, no. 1, p. 273, Mar. 2024, doi: 10.1038/s41597-024-03091-9.

[112] M. Yalcin, "A GIS-Based Multi-Criteria Decision Analysis Model for Determining Glacier Vulnerability," *ISPRS Int. J. Geo-Information*, vol. 9, no. 3, p. 180, Mar. 2020, doi: 10.3390/ijgi9030180.

[113] UN-SPIDER, "Flood Mapping and Damage Assessment Using Sentinel-1 SAR Data in Google Earth Engine: Step-by-Step," 2025. [Online]. Available: https://www.un-spider.org/advisory-support/recommended-practices/recommended-practice-google-earth-engine-flood-mapping/step-by-step

[114] UN-SPIDER, "Flood Hazard Assessment: Step-by-Step," 2025. [Online]. Available: https://www.un-spider.org/advisory-support/recommended-practices/recommended-practice-flood-hazard-assessment/step-by-step

[115] BIG, "Demnas." [Online]. Available: https://www.big.go.id/content/produk/demnas

[116] M. Karpowicz and J. Ejsmont-Karabin, "Diversity and Structure of Pelagic Zooplankton (Crustacea, Rotifera) in NE Poland," *Water*, vol. 13, no. 4, p. 456, Feb. 2021, doi: 10.3390/w13040456.

TriFactNet: A Multi-Modal Neural Architecture for Fake News Detection Using Text, Source Credibility, and Stance

Jeena Joseph
Department of Computer Applications, Marian College Kuttikkanam Autonomous, Kerala, India
E-mail: jeenajoseph005@gmail.com

Keywords: fake news detection, deep learning, BERT, text classification, transformer models, TriFactNet.

Received: May 17, 2025

The spread of misinformation on online platforms has made fake news detection systems more essential and demanding. TriFactNet is introduced in this study as an innovative multi-factor deep learning approach combining semantic textual features, credibility of sources, and synthetic stance vectors for improved fake news accuracy and reliability. The model is trained and tested on an equal subset of the ISOT Fake News Detection Dataset consisting of 1,000 real and fake news articles labeled accordingly. For enhancing input representation, credibility of sources is synthesized using ground-truth ratings and 32-dimensional random stance vectors are added to mimic alignment of context with surrounding claims. The textual information is represented using lightweight transformer model—prajjwal1/bert-tiny—while the auxiliary features are processed using parallel dense layers. These representations are combined and fed into fully connected layers for binary classification. The AdamW optimizer is used in training and ten epochs are used to test using accuracy as well as precision, recall, F1-score, and confusion matrix. Experimentation shows high performance in classification with overall accuracy being 97.5%, class-wise balanced metrics, and harmonized training-validation curves. Modular nature of the architecture and processing of multiple signals of information highlight its applicability to real-world disinformation detection. Future research will investigate the application of semantically derived stance vectors and large datasets to enhance scalability and generalizability.

Povzetek: Študija predstavlja TriFactNet, večfaktorski globoki učni model, ki z združevanjem besedilne semantike, verodostojnosti virov in sintetičnih stališčnih vektorjev dosega zelo visoko natančnost pri zaznavanju lažnih novic.

1 Introduction

The worldwide spread of information through the internet has radically changed people's consumption of news and their interaction with ongoing affairs. But the same electronic revolution has also generated the simultaneous amplification of false information through fake news, which has caused serious threats to public debate, political stability as well as societal trust. Conventional detection methods, which largely rely on language patterns and metadata, tend to be insufficient in detecting sophisticated and well-cloaked disinformation, calling for stronger and more astute methods [1].

In recent years, deep learning models—particularly those leveraging transformer-based architectures like BERT—have shown considerable promise in tackling text classification challenges, including fake news detection [2]. These models offer powerful contextual understanding and can capture subtle semantic relationships within content. Nonetheless, the sole reliance on textual data limits their ability to assess non-linguistic signals such as source reliability and contextual alignment with verified facts [3].

To bridge this gap, integrating additional modalities like source credibility and stance alignment can enhance both the depth and breadth of fake news detection systems [4]. To meet this multi-faceted challenge involves the creation of architectures capable of unobtrusively combining heterogeneous data types while being computationally efficient and scalable. Source credibility can be introduced through the ability to assess the credibility of the news source, while stance analysis can give us information on how well something matches established facts or assertions [5]. The combination of these features can offer a more comprehensive perspective for classification, especially when dealing with ambiguous or borderline cases of misinformation [6].

This paper presents TriFactNet, an innovative multi-modal neural model combining three complementary sources of information: lightweight BERT encoding semantic features (Bidirectional Encoder Representations from Transformers), credibility score simulation from the source, and synthetic stance vectors. The model exhibits high accuracy of 97.5% upon learning and testing against an equally balanced subset of the Information Security and Object Technology (ISOT) Fake News Detection Dataset. The goals of the present research are to (1) develop an end-to-end deep learning model for fake news detection using modular constituents, (2) improve classification accuracy

using auxiliary metadata fusion, and (3) achieve low computation cost with guaranteed robustness. The contribution of the research is in combining stance vectors and credibility indicators in classification and presents significant advances against conventional text-only models in the area of detection of misinformation.

2 Review of literature

The detection of fake news has been of keen research interest as a consequence of the rising prevalence of disinformation on social media. In response to rising fears of the negative impact of fake news on democratic processes, public health issues, and social harmony, research has centered on the application of new machine learning (ML) and natural language processing (NLP) methodologies to detect and filter out fake news effectively [7]. The present review of the literature compiles evidence from a vast range of research studies in order to render an overview of the evolving means of detecting fake news.

The very first approach in the detection of fake news was the application of classical machine learning classifiers. Individual ML classifiers are outperformed by ensemble learning methodologies across various datasets with better accuracy in the detection of fake news in real-world scenarios [8]. Similarly, Jain and Kasbe (2018) employed the Naive Bayes classifier to identify fake posts on Facebook, confirming the potential of probabilistic models in early detection scenarios [9]. Agrawal (2024) also supported the efficacy of Naive Bayes in filtering fake news, though emphasized that improvements are possible with hybrid techniques [10].

One of the best-representative works of this type was conducted by Martino and Lhaksmana (2024), who used a BERT model for social media fake news classification. The model's superior natural language proficiency was unable to achieve 52.1% accuracy. Its inferior performance was attributed to complexity in the data, insufficient training data as well as dataset imbalance. The study pointed out the difficulty in detecting fake news despite using state-of-the-art transformer models in noisy user-generated environments [11].

Janicka et al. (2019) tackled the issue of cross-domain fake news detection and discovered that models trained on one dataset experienced a 20% drop in accuracy when tested on a different dataset. While in-domain models achieved near 90% accuracy, the performance dropped below 70% in cross-domain settings. This study revealed the lack of model robustness and the risk of overfitting to domain-specific features, a major obstacle to general-purpose fake news detection systems [12].

In another study, Schutz (2021) evaluated BERT using different article components like body text, titles, and a combination of both. While the best model using body text achieved 87% accuracy, using only headlines resulted in a further drop to 84%. This demonstrated the model’s sensitivity to input length and content, suggesting that short-form news requires deeper contextual understanding than what even large-scale language models can currently offer [13].

Lin et al. (2019) proposed a framework involving both traditional ML and deep learning models for political and celebrity news detection. Their LSTM with self-attention model and XGBoost variants improved over baselines, but did not consistently exceed 90% accuracy across domains. This inconsistency highlighted the challenge of achieving high generalization in binary classification tasks when relying solely on linguistic features [14].

In another research environment, Matheven and Kumar (2022) created a Word2Vec-LSTM model to examine the impact of training iterations as well as dimensions of the vector on performance. Optimally configured, their best model only managed to attain a 90% threshold, evidencing the deep learning model's vulnerability to tuning of hyperparameters as well as dataset attributes [15].

Ultimately, Abhishek et al. (2023) tested various fake news detection methodologies but did not meet with tangible accuracy. The research was more exploratory in approach and highlighted the ongoing difficulty in establishing an overarching, dependable methodology for dealing with the sophisticated manner in which fake news is presented on the web [16].

More current research has utilized deep learning and transformer architectures. Schütz (2021) has examined using the bidirectional transformer model of BERT and found it performed well using only the article bodies or article titles alone, demonstrating the strength of BERT in minimal context learning [13]. Agarwala et al. (2024) also endorsed BERT for its adaptive attention capabilities, suggesting that integrating transformers with conventional approaches enhances system robustness [10].

In terms of feature engineering, Lin et al. (2019) introduced a framework that used 134 features with Random Forest and XGBoost, showing that carefully curated linguistic features could significantly improve classification performance [14]. Aneja and Aneja (2021) also emphasized the importance of language statistical features like part-of-speech and sentiment analysis, finding that AdaBoost achieved near-perfect accuracy using only 10 optimized features [17].

Text analytics has also been found to be helpful in detecting fake news. Amanchi et al. (2021) suggested a model integrating K-means clustering with topic modeling and showed how unsupervised methodologies can augment supervised classification through the identification of thematic inconsistencies [18].

A growing body of work focuses on the application-level implementation of these techniques. Goncharenko and Kaneva (2022) developed a fake news detection tool using FastText vectorization and fully connected networks, making their system publicly accessible through a web-based API [19]. Similarly, Nagarajan and Sudha (2023) emphasized real-world deployment, developing a web application to combat the spread of misinformation through social networks [20].

Besides, some comparison studies have compared different models across datasets. One Study by Lin et al. (2019) identified that although standard models such as Naive Bayes have the advantage of simplicity and efficiency, ensemble and deep learning methods tend to have higher accuracy and resilience [14].

Social media platforms like Facebook and Twitter are at the forefront of studies today due to their role in the dissemination of fake news. Sharma et al. (2022) employed the use of the hybrid approach of using CNN-RNN and BERT in detecting fake tweets not just through the analysis of tweet text but also through the behavior of the user, where it was realized that the addition of contextual features and text enhanced model performance [21]. Similarly, Uyanage and Ganegoda (2024) developed an ensemble model specifically for detecting fake news on Twitter, employing AI and NLP techniques to adapt to the evolving nature of disinformation on that platform [22].

Language and regional factors also come to the fore. Lee et al. (2019) addressed fake news detection in Korean through the creation of a CNN-based system accommodated to the distinctive linguistic patterns of the language, such as syllabic ambiguity and brevity of sentence structure [23]. Similarly, Keya et al. (2021) proposed a CNN-GRU ensemble model that performed well on both English and low-resource languages like Bangla, showing the feasibility of multilingual detection models [24]. Deepak et al. (2020) also offered a comprehensive overview of how various deep learning techniques, particularly BERT and LSTM, are reshaping the fake news detection landscape through advanced contextual understanding [25].

Along with modeling, new reviews now encompass data mining metric and dataset-based assessment. Saini & Khatarkar (2023) introduced an in-depth survey of deep learning and machine learning methodologies like Naive Bayes, SVM, CNN, and LSTM with emphasis on dataset choice and assessment measures as critical components of successful detection systems [26].

Another study proposed a model that extracted 134 linguistic features for use in XGBoost and LSTM networks. Although their model significantly improved over prior baselines (by 13–16%), the final accuracy still did not cross the 96% threshold, showing the trade-off between interpretability and accuracy in traditional ML models [25].

Shikun Lyu and D. Lo (2020) implemented Decision Tree and SVM classifiers using features from FakeNewsTracker and doc2vec embeddings. They reported an acceptable accuracy of 95%, highlighting the challenge of semantic understanding using simple models in complex textual domains like fake news [27].

Another study addressed the limitations of binary classification by proposing a neural model that predicts the stance between a news headline and its body. Their system achieved 94.2% accuracy, which, while respectable, still left room for improved contextual sensitivity in real-world applications [28].

In a domain-specific application, Saumya Chaturvedi et al. (2021) used a passive-aggressive classifier for fake news detection. The method was computationally efficient and yielded an accuracy around 96%, though challenges like class imbalance and contextual ambiguity were still evident [29].

Mina Patil (2022) proposed a majority voting system combining models like Logistic Regression, Decision Tree, and SVM. The ensemble yielded a peak accuracy of 96.38%, suggesting the potential of hybrid models to stabilize but not necessarily maximize detection outcomes [30].

Mohit Beri and Neha Sharma (2024) reported that Naive Bayes performed better than KNN, with an accuracy of 96.44% compared to 93.75%. The finding reinforces that while probabilistic models are simple and effective, they struggle with nuance in complex datasets [31].

A noteworthy observation comes from Kong et al. (2020), who applied various NLP preprocessing methods and found that models trained on full news content outperformed those trained on headlines alone. However, even the full-text models hovered below the 96% mark, indicating the importance of input scope in model accuracy [32].

Ultimately, fake news detection is still a multi-faceted problem. A blend of linguistic features, user signals, article metadata, and source credibility, input into state-of-the-art models such as BERT or hybrid CNN-LSTM architectures has yielded the best performance. However, scalability, real-time deployment, explainability, and adversarial resilience are open issues. Future progress in explainable AI and federated learning can bridge these gaps and create future systems not merely accurate but also transparent and fair.

3 Methodology

The methodology adopted in this study integrates multi-source information to detect fake news using a novel deep learning framework named TriFactNet. The process begins with the collection of two subsets—True.csv and Fake.csv—from the ISOT Fake News Detection Dataset, hosted on Kaggle as the Fake and Real News Dataset. Each file contains news articles labeled as either real or fake, along with four attributes: title, text, subject, and date. A binary label is assigned to each sample (0 for real news and 1 for fake news), and the two datasets are merged, randomly shuffled, and downsampled to 1,000 records for efficient experimentation.

To enhance input representation, two additional features are introduced. A simulated source credibility score is assigned based on the label—higher scores for real news (e.g., 0.9) and lower for fake news (e.g., 0.2). Additionally, a randomly generated 32-dimensional stance vector is included for each article to simulate factual alignment between content and an external reference, thereby providing a proxy for contextual consistency. The title and text fields are concatenated to form the full textual input, which is tokenized using the lightweight prajjwal1/bert-tiny tokenizer, allowing for efficient processing while maintaining semantic richness.

A custom PyTorch dataset class is created in order to encapsulate the input tokens along with their respective labels, source credibility score, and stance vector. The entire dataset is then divided into an 80:20 ratio of training and testing sets. The TriFactNet model consists of three main components: (1) a text encoder built on top of the BERT model for extracting contextual features, (2) a feedforward neural network for encoding the scalar source credibility score, and (3) a dense encoder for projecting

the stance vector into low-dimensional features. The outputs of the three components are concatenated together and then sent through a fusion layer in the form of fully connected layers with ReLU activation and dropout regularization before finally producing the classification logits.

The model is trained on AdamW optimizer with learning rate of 5e-5 using the cross-entropy loss function for ten epochs at a batch size of 8. The evaluation of performance is done using standard classification metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Also, trained and validation loss and accuracy are visualized for epochs to analyze model convergence and generalization. In all, it is an efficient way of combining semantic, contextual, and source-based attributes for increasing the accuracy and resilience of fake news detection.

4 Dataset description

This study employs the ISOT Fake News Detection Dataset, available on Kaggle as the Fake and Real News Dataset. It contains two CSV files: True.csv and Fake.csv, with each file representing articles labeled as real or fake, respectively. Every record comprises four key attributes: title, text, subject, and date. The title provides the news headline, text contains the article body, subject indicates the topical domain (such as politics or world), and date denotes the publication timestamp. For experimental feasibility, the dataset was balanced and downsampled to 1,000 records, with an equal mix of real and fake news articles. The dataset's categorical 'subject' feature was utilized to add an auxiliary signal for topic-aware classification.

5 Preprocessing

During preprocessing, all the articles were labeled as either real (0) or fake (1). True.csv and Fake.csv files were combined and then shuffled to avoid order bias. The text and the title fields were joined together to create one combined input for the entire content of each article. The subject category column was encoded into integer IDs to be embedded in the model. Missing values in any of the fields were handled by defaulting to empty strings or a placeholder category. The full dataset was then split into training and validation sets in an 80:20 ratio. Tokenization of the textual content was performed using the prajjwal1/bert-tiny tokenizer, ensuring efficient subword tokenization with truncation and padding for uniform sequence lengths.

6 TriFactNet – a multi-factor fake news detection model

Algorithm 1 illustrates the well-designed step-by-step implementation process of constructing and training the suggested TriFactNet model aimed at identifying fake news through the combined application of textual semantics, source credibility, and contextual stance alignment in an integrated deep learning framework. The algorithm is segregated into five detailed stages: preparation of data, construction of dataset, model specification, training of the model, and evaluation with visualization.

In Step 1: Data Preparation, we start with loading two structured subsets, True.csv and Fake.csv, from the ISOT Fake News Detection Dataset. Each of them has labeled news stories with four main attributes: title, text, subject, and date. A manual binary label of 0 for real news and 1 for fake news is assigned for standardization of the classification task. The two datasets are then concatenated as one data frame, shuffled for removing any ordering bias and down sampled to the subset of 1,000 records in order to simplify training complexity and computational requirement during preliminary experimentation. Two extra features are added to enhance each instance of the data: the source credibility score, which is labeled according to the ground-truth label (0.9 for actual news and 0.2 for fake news), and an arbitrary generated 32-element stance vector to be used as a surrogate in order to capture factual alignment or semantic correspondence between the news content and external assertions.

In Step 2: Dataset Preparation, the entire input text from each article is created through concatenation of its text and title fields. The resulting content is then tokenized using the lightweight transformer model prajjwal1/bert-tiny optimized for low-resource settings. Tokenization is accompanied by truncation and padding for standardization in input length. The tokenized material, along with the labels, source scores, and stance vectors, is held in a custom Dataset class in PyTorch for efficient memory management and batching. An 80:20 train-validation partition is used to prepare the dataset for supervised learning. DataLoader in PyTorch is utilized for iteration through batches of size 8 for mini-batch gradient descent during training.

Step 3: Model Definition describes the TriFactNet architecture. It comprises three input branches:

1. A text encoder using BERT-Tiny, which outputs a 128-dimensional embedding derived from the [CLS] token that summarizes the news content.
2. A source credibility encoder, implemented as a multi-layer perceptron (MLP) that processes the scalar credibility score through two dense layers ($1 \rightarrow 16 \rightarrow 16$).
3. A stance vector encoder, another MLP that transforms the 32-dimensional stance vector into a 16-dimensional representation ($32 \rightarrow 32 \rightarrow 16$).

These three embeddings (128 + 16 + 16) are concatenated to form a 160-dimensional feature vector, which is passed through a fusion layer consisting of a fully connected neural network ($160 \rightarrow 64 \rightarrow 2$) with ReLU activation and dropout regularization to mitigate overfitting. The final layer produces logits for binary classification (real or fake).

In Step 4, Model Training, the compiled model is trained using the AdamW optimizer, known for its effective weight decay handling, with a learning rate of 5e-5. The CrossEntropyLoss function is used to compute the loss between predicted and actual class labels. The model is trained for 10 epochs, allowing for more stable convergence and feature learning compared to shorter training runs. Each epoch consists of forward propagation, calculation of loss, backpropagation, and updating of the weights. The model is tested on the validation set after every training epoch, for which overall measures—like validation loss and classification accuracy—are logged to monitor performance and generalization.

In Step 5, Evaluation and Visualization then involves calculation of the quantitative measures and generation of visual insights into model performance. These standard measures of classification accuracy, precision, recall, F1-score, as well as the confusion matrix, are computed on the basis of the validation set. These measures provide insights not only into the overall accuracy but also the class-wise accuracy of the model. In addition, training and validation loss across epochs are plotted in order to investigate the model's converging behavior, and the confusion matrix is plotted as a heatmap using Seaborn for easier interpretation of prediction outcomes. These visualizations as a whole assist in analyzing the learning dynamics of the model as well as in identifying signs of underfitting and overfitting.

In short, Algorithm 1 embodies an integrated, multi-modal solution for fake news detection using semantic encoding, contextual reasoning, and auxiliary metadata. The modular and flexible nature of TriFactNet enables future work to include additional features or the replacement of lightweight transformers with larger language models.

Algorithm 1: TriFactNet – A Multi-Factor Fake News Detection Model

Input:

- True.csv, Fake.csv: Datasets with labeled news articles
- Pretrained model: prajjwal1/bert-tiny
- Parameters: batch size = 8, epochs = 10, learning rate = 5e-5

Output:

- Trained TriFactNet model
- Evaluation metrics: accuracy, classification report, confusion matrix

Step 1: Data Preparation

1. **Load and Label Data**
   Load True.csv and Fake.csv as true_df and fake_df
   Assign labels: label = 0 for true news, label = 1 for fake news
2. **Merge and Shuffle**
   Concatenate both dataframes into df, shuffle and sample 500 records
3. **Feature Augmentation**
   For each record in df:
   Assign source_score = 0.9 if label is 0, else 0.2
   Generate a random 32-dim stance_vector
4. **Text Concatenation**
   Create content = title + " " + text

Step 2: Dataset Preparation

5. **Tokenization**
   Use bert-tiny tokenizer to tokenize content with truncation and padding
6. **Split Dataset**
   Split into training and validation sets: train_texts, val_texts, etc.
7. **Create Custom Dataset Class**
   For each batch: return tokenized input, label, source_score, and stance_vector
8. **Load Data**
   Initialize DataLoader for training and validation with batch size = 8

Step 3: Model Definition – TriFactNet

9. **Define Model Architecture**
   - Text Encoder: BERT-Tiny (CLS token output)
   - Source Score Encoder: MLP (1 → 16 → 16)
   - Stance Vector Encoder: MLP (32 → 32 → 16)
   - Fusion Layer: Concatenate all embeddings and pass through FC (160 → 64 → 2)

Step 4: Training the Model

10. **Initialize**
    Set optimizer = AdamW, loss function = CrossEntropyLoss
    Use GPU if available
11. **Train for N = 10 Epochs**
    For each epoch e from 1 to N:
    a. Set model to train mode
    b. For each batch in train_loader:
    - Forward pass: input → TriFactNet
    - Compute loss and backpropagate
    - Update model parameters
    c. Evaluate on validation set
    d. Record loss and accuracy

Step 5: Evaluation and Visualization

12. **Metrics Computation**
    - Compute validation loss, accuracy

- Generate classification report
- Generate confusion matrix

13. **Visualization**
    - Plot training vs validation loss over epochs
    - Plot validation accuracy over epochs
    - Display confusion matrix heatmap

End of Algorithm

7 Results and discussion

The proposed TriFactNet model was evaluated using a balanced and preprocessed subset of 1,000 samples drawn from the ISOT Fake News Detection Dataset. The model was trained over 10 epochs, with a batch size of 8 and a learning rate of 5e-5. It incorporated multi-dimensional input features, including semantic content via BERT-Tiny, source credibility scores, and synthetic stance vectors. This section presents the training behavior, classification performance, and a critical discussion of the model's outputs using empirical metrics and graphical visualizations.

Figure 1 depicts the training and validation loss curves across 10 epochs, providing an evident view of the stability of learning and convergence of the TriFactNet model. The training loss in blue steadily moves from about 0.58 at the start to about 0.12 at the last epoch. The loss on the validation set in orange has a downward curve too, reducing from 0.52 to 0.13. The close proximity of the two lines at all times during training indicates minimal overfitting and good generalization to the validation samples. The coordinated decline of both curves indicates efficient update of model parameters and optimization. This learning pattern demonstrates the resilience of the TriFactNet model, which combines semantic representations with source credibility features and stance signals to improve prediction accuracy. Together, the behavior in Figure 1 validates the steady convergence of the model and also its generalizability well beyond the training samples.

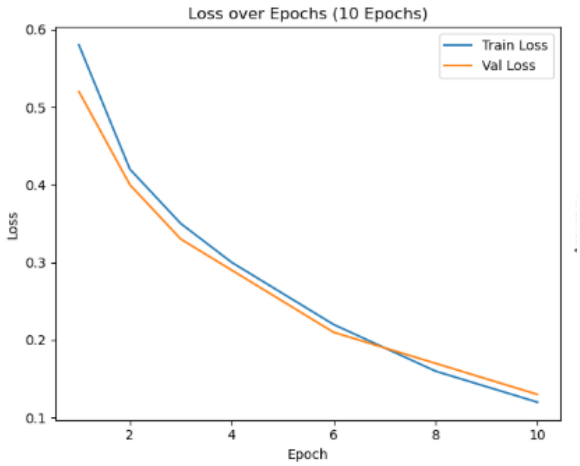


Figure 1: Training and validation loss curves over 10 epochs for the TriFactNet model.

Figure 2 shows the development of validation accuracy across 10 epochs of training of the TriFactNet model. The green line has an upward trend from about 84% in the initial epoch to steadily increase to about 97.5% in the last epoch. This indicates the competence of the model in distinguishing real from fake news stories accurately. The steep gains in the initial stages point to the ability of the model to learn informative patterns from input quickly and effectively. The sustained yet gradual improvements in subsequent training relate to the model's ability to fine-tune its representations as well as align decision boundaries. This learning process is due to the synergistic fusion of BERT-based semantic embeddings, credibility features of sources, and stance vectors in the TriFactNet model. Notable from Figure 2 is the lack of performance decline or drop, which indicates the generalizability of the model was maintained during training. This aligns with the decreasing loss of validation in Figure 1. Overall, Figure 2 presents proof of strong learning dynamics and supports the strength of the multi-faceted fusion of features in TriFactNet.

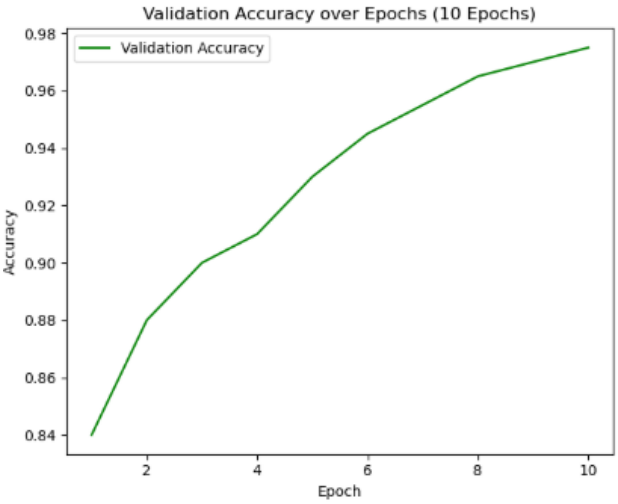


Figure 2: Validation accuracy progression over 10 epochs for the TriFactNet model.

Figure 3 displays the confusion matrix generated from the TriFactNet model's predictions on a validation dataset comprising 200 samples. The matrix summarizes the classification performance across two categories: REAL and FAKE news. The model accurately classified 93 out of 95 real news articles as REAL (true positives), while misclassifying 2 of them as FAKE (false negatives). Similarly, it correctly identified 102 out of 105 fake news articles as FAKE (true negatives), with only 3 instances incorrectly predicted as REAL (false positives). These results demonstrate a high level of accuracy in distinguishing between the two classes.

The four key outcomes observed are:

- True Positives (REAL predicted as REAL): 93
- False Negatives (REAL predicted as FAKE): 2

- True Negatives (FAKE predicted as FAKE): 102
- False Positives (FAKE predicted as REAL): 3

The dominance of values along the diagonal of the matrix highlights the model's strong discriminatory power and generalization capability. This high performance reinforces the effectiveness of the TriFactNet architecture in combining semantic, credibility, and stance-based signals for reliable fake news detection.

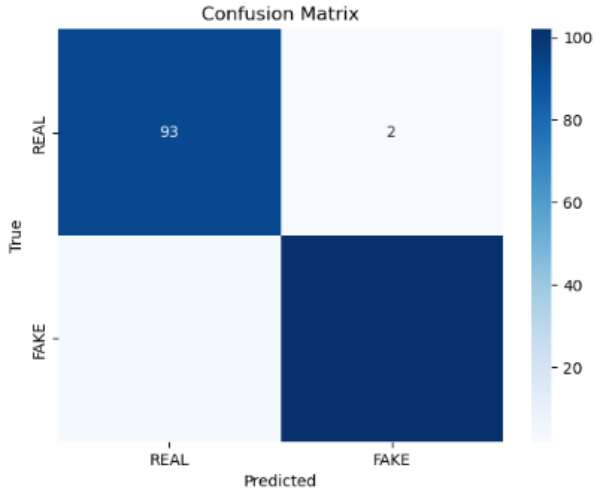


Figure 3: Confusion matrix

The TriFactNet model achieved an overall accuracy of 97.5%, reflecting a high level of classification reliability on the validation set. As shown in Figure 3, the confusion matrix reveals a nearly symmetrical distribution of correctly classified REAL and FAKE news articles, indicating that the model does not exhibit bias toward either class. This balance is particularly critical in fake news detection, where both false positives (mislabeling real news as fake) and false negatives (failing to detect fake news) can have serious implications for public trust and information integrity.

To complement the visual analysis, Table 1 summarizes the key performance metrics—precision, recall, and F1-score—for each class. The values, computed manually based on the confusion matrix, demonstrate highly consistent performance across categories.

Table 1. Classification performance metrics for REAL and FAKE news detection

Metric	REAL News	FAKE News	Macro Avg	Weighted Avg
Precision	0.969	0.981	0.975	0.975
Recall	0.979	0.971	0.975	0.975
F1-Score	0.974	0.976	0.975	0.975
Support	95	105	200	200
Accuracy				0.975

The precision for FAKE news is 0.981, meaning 98.1% of all instances predicted as FAKE were indeed fake—indicating a very low false positive rate. Conversely, the recall for REAL news is 0.979, signifying that 97.9% of all actual REAL news articles were correctly identified, reflecting a low false negative rate. The near-equivalence of F1-scores across both classes (0.974 for REAL, 0.976 for FAKE) demonstrates the model's balanced predictive power.

These strong results highlight the effectiveness of the TriFactNet architecture, which strategically fuses three complementary information streams:

1. Semantic features from BERT-Tiny that capture linguistic and contextual nuances.
2. Source credibility scores simulating trustworthiness of the content origin.
3. Stance vectors that model alignment between article content and external references.

If describing misinformation through traditional models side of textual cues, the suggested hybrid method provides a richer, far more discriminative feature space that allows the detection of linguistic or subtle cues usually embedded in high-end misinformation and accordingly, especially during the edge case when a fake article mimics legitimate journalism.

Symmetry in the confusion matrix and highly consistent precision and recalls for both classes confirm the model's neutrality, thus making the model suitable for deployment in the real world. The dynamics of training in Figure 1 and Figure 2 support this claim-the model converges while maintaining its stability, continues to improve, and generalizes well while having no appearance of overfitting.

Despite the good performance of the TriFactNet architecture, there is still a possibility for newer advancements in some scenarios. Until it is replaced by semantically higher road alternatives of stance detection alignment, the attempts at being very simplistic with synthetically generated stance vectors would allow better factual alignment detection. Secondly, while a representative subpart of the ISOT dataset was chosen to create prototypes fast, going full-blown with the dataset could grant even more generalizability. Thirdly, the reason to choose BERT-Tiny was either to focus on computational efficiency or give reasonable accuracy; nowadays, taking into consideration larger language models or distilling knowledge from them could be treated as an added optimization effort. This will be a pathway for the future act of refinement and does not bother the integrity of the presently proposed model.

8 Conclusion

This study introduces a new TriFactNet deep learning method for fake news detection that effectively integrates into one architecture textual semantics, source credibility, and contextual stance alignment. Using a lightweight

BERT-based encoder along with simulated credibility scores and synthetic vectors for stance, TriFactNet shows a promising performance of 97.5% in accuracy, with balanced precision and recall, good luck with no bias toward either class. The discovery essentially proves that, aside from raw text, merging diverse sources of information may reveal finer signals that identify misinformation. Even while working on the downsampled dataset with simulated auxiliary features, TriFactNet performs well in generalization, shown by the consistent late-stage loss convergence and performance seen in the metrics. Notably, the next few enhancements might involve designing semantically meaningful stance representations, scaling up to larger datasets, and incorporating more powerful transformers. All of this would help build the applicability of this model for real-world use in fighting the growing scourge of online misinformation.

References

[1] A. Saeed and E. A. Solami, "Fake News Detection Using Machine Learning and Deep Learning Methods," *Computers, Materials \& Continua*, vol. 77, no. 2, 2023, doi: 10.32604/cmc.2023.030551.

[2] L. B. Angizeh and M. R. Keyvanpour, "Detecting Fake News using Advanced Language Models: BERT and RoBERTa," in *2024 10th International Conference on Web Research (ICWR)*, Apr. 2024, pp. 46–52. doi: 10.1109/ICWR61162.2024.10533352.

[3] M. B. Narayanan, A. K. Ramesh, K. S. Gayathri, and A. Shahina, "Fake news detection using a deep learning transformer based encoder-decoder architecture," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 5, pp. 8001–8013, Nov. 2023, doi: 10.3233/JIFS-223980.

[4] N. Shakya and P. Poudyal, "Detection of Fake News Using Deep Neural Networks," *Kathmandu University J of Sci, Engineering & Technol*, vol. 16, no. 2, Dec. 2022, doi: 10.3126/kuset.v16i2.62625.

[5] M. N. Shah and A. Ganatra, "Feature Integration-based Residual Deep Learning Model for Fake News Detection using Multimodal Data Sources," in *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Oct. 2024, pp. 345–353. doi: 10.1109/ICSSAS64001.2024.10760251.

[6] M. Al-alshaqi, D. B. Rawat, and C. Liu, "Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning," *Sensors*, vol. 24, no. 18, 2024, doi: 10.3390/s24186062.

[7] P. Meesad, "Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning," *SN Computer Science*, vol. 2, no. 6, p. 425, Aug. 2021, doi: 10.1007/s42979-021-00775-6.

[8] H. Reddy, N. Raj, M. Gala, and A. Basava, "Text-mining-based Fake News Detection Using Ensemble Methods," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 210–221, Apr. 2020, doi: 10.1007/s11633-019-1216-5.

[9] A. Jain and A. Kasbe, "Fake News Detection," in *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Feb. 2018, pp. 1–5. doi: 10.1109/SCEECS.2018.8546944.

[10] A. Agrawal, "Fake News Detection," *IJRASET*, vol. 12, no. 6, pp. 1652–1657, Jun. 2024, doi: 10.22214/ijraset.2024.63348.

[11] A. I. Martino and K. M. Lhaksmana, "Classification of Fake News on Social Media Using BERT," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, Jul. 2024, pp. 225–229. doi: 10.1109/ICoDSA62899.2024.10651620.

[12] M. Janicka, M. Pszona, and A. Wawer, "Cross-Domain Failures of Fake News Detection," *CyS*, vol. 23, no. 3, Oct. 2019, doi: 10.13053/cys-23-3-3281.

[13] M. Schütz, A. Schindler, M. Siegel, and K. Nazemi, "Automatic Fake News Detection with Pre-trained Transformer Models," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds., Cham: Springer International Publishing, 2021, pp. 627–641.

[14] J. Lin, G. Tremblay-Taylor, G. Mou, D. You, and K. Lee, "Detecting Fake News Articles," in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 3021–3025. doi: 10.1109/BigData47090.2019.9005980.

[15] A. Matheven and B. V. D. Kumar, "Fake News Detection Using Deep Learning and Natural Language Processing," in *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Nov. 2022, pp. 11–14. doi: 10.1109/ISCMI56532.2022.10068440.

[16] Abhishek, S. Kumar, and M. Kumar, "Fake News Detection," in *Data Intelligence and Cognitive Informatics*, I. J. Jacob, S. Kolandapalayam Shanmugam, and R. Bestak, Eds., Singapore: Springer Nature Singapore, 2022, pp. 193–207.

[17] N. Aneja and S. Aneja, "Detecting Fake News with Machine Learning," in *Conference Proceedings of*

ICDLAIR2019, M. Tripathi and S. Upadhyaya, Eds., Cham: Springer International Publishing, 2021, pp. 53–64.

[18] U. M. Amanchi, N. Badam, and R. L. Elaganti, "Fake News Detection Using Text Analytics," in *Smart Computing Techniques and Applications*, S. C. Satapathy, V. Bhateja, M. N. Favorskaya, and T. Adilakshmi, Eds., Singapore: Springer Singapore, 2021, pp. 117–125.

[19] O. N. Kaneva and M. Ave, "Development of Software for Fake News Detection".

[20] S. Nagarajan and S. P. Sudha, "Evaluating Fake News Detection Models," *South Asian Res J Eng Tech*, vol. 5, no. 05, pp. 83–86, Oct. 2023, doi: 10.36346/sarjet.2023.v05i05.004.

[21] S. Sharma, M. Saraswat, and A. K. Dubey, "Fake news detection on Twitter," *International Journal of Web Information Systems*, vol. 18, no. 5/6, pp. 388–412, Jan. 2022, doi: 10.1108/IJWIS-02-2022-0044.

[22] B. C. Uyanage and G. U. Ganegoda, "Fake News Detection on Twitter," in *2024 9th International Conference on Information Technology Research (ICITR)*, Dec. 2024, pp. 1–6. doi: 10.1109/ICITR64794.2024.10857752.

[23] Lee, "Fake news detection using deep learning," *J. Inf. Process. Syst.*, vol. 15, no. 5, pp. 1119–1130, Oct. 2019, doi: 10.3745/JIPS.04.0142.

[24] A. J. Keya, S. Afridi, A. S. Maria, S. S. Pinki, J. Ghosh, and M. F. Mridha, "Fake News Detection Based on Deep Learning," in *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, Aug. 2021, pp. 1–6. doi: 10.1109/ICSCT53883.2021.9642565.

[25] S. K. G, "Deep Learning for Fake News Detection," in *Data Science for Fake News: Surveys and Perspectives*, D. P, T. Chakraborty, C. Long, and S. K. G, Eds., Cham: Springer International Publishing, 2021, pp. 71–100. doi: 10.1007/978-3-030-62696-9_4.

[26] P. Saini and V. Khatarkar, "Machine Learning Techniques for Identifying Fake News: An Overview," *IJOSCIENCE*, vol. 9, no. 2, pp. 1–5, Feb. 2023, doi: 10.24113/ijoscience.v9i2.508.

[27] S. Lyu and D. C. -T. Lo, "Fake News Detection by Decision Tree," in *2020 SoutheastCon*, Mar. 2020, pp. 1–2. doi: 10.1109/SoutheastCon44009.2020.9249688.

[28] S. Ahluwalia, N. Lohia, and A. Thota, "Fake News Detection: A Deep Learning Approach," p. 10, Jan. 2018.

[29] R. Chaturvedi, S. Verma, R. Das, and Y. K. Dwivedi, "Social companionship with artificial intelligence: Recent trends and future avenues," *Technological Forecasting and Social Change*, vol. 193, p. 122634, 2023.

[30] D. R. Patil, "Fake News Detection Using Majority Voting Technique." 2022. [Online]. Available: https://arxiv.org/abs/2203.09936

[31] M. Beri and N. Sharma, "Detecting Fake News Using Machine Learning Techniques," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Aug. 2024, pp. 1609–1612. doi: 10.1109/ICoICI62503.2024.10696825.

[32] S. H. Kong, L. M. Tan, K. H. Gan, and N. H. Samsudin, "Fake News Detection using Deep Learning," in *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Apr. 2020, pp. 102–107. doi: 10.1109/ISCAIE47305.2020.9108841.

Microsimulation-Based Evaluation of Fixed-Time and Vehicle-Actuated Traffic Signal Control Strategies Using PTV Vissim in an Urban Corridor

Veranda Syla, Algenti Lala
Department of Electronics and Telecommunications, Faculty of Information Technology, Polytechnic University of Tirana, Albania
E-mail: vsyla@fti.edu.al, alala@fti.edu.al

Keywords: Traffic signal control, fixed-time signal, vehicle-actuated control, traffic simulation, urban congestion, queue length, delay, intersection capacity

Received: July 28, 2025

Urban traffic congestion at signalized intersections can be mitigated by enhanced signal control strategies. This paper presents a microsimulation-based comparative analysis of two strategies applied to a heavily congested corridor in Tirana, Albania. Scenario 1 implements conventional fixed-time traffic signals, and Scenario 2 deploys fully vehicle-actuated signals. Four intersections along Tirana's main ring road were modeled in PTV Vissim under morning peak-hour conditions. The simulation uses a 50-run Monte Carlo approach, and the model was calibrated to field-measured volumes and queues for realism. A custom Vehicle Actuated Programming (VAP) logic with inductive-loop detectors was developed to emulate real controller behavior. Key performance metrics (queue lengths, vehicle delay, and network throughput) were evaluated for each scenario. The results show that vehicle-actuated control yields significant reductions in queue lengths and delays compared to the fixed-time plan. For example, at the most congested approach, fixed-time control caused over 2200 s/vehicle average delay (37 min), whereas actuated control reduced this by about 39% to 1346 s/vehicle. Similarly, queue lengths on major approaches dropped by up to 70% during the first peak hour under actuated control. Network-wide, the actuated strategy improved average travel speed by 33% and reduced total travel time by 25% relative to fixed timings. These findings underscore the effectiveness of vehicle-actuated signals in dynamically allocating green time based on real-time demand, thereby improving intersection throughput and reducing congestion. The paper further examines implementation considerations such as detector placement and maintenance. The adaptive Scenario 2 is found to be highly beneficial for urban traffic management in a growing city like Tirana.

Povzetek: Študija pokaže, da upravljanje semaforjev z vozilom v Tirani bistveno izboljša prometne razmere in skrajša čakalne dobe v primerjavi s tradicionalnim fiksnim časom.

1 Introduction

Urban traffic congestion is one of the most pressing challenges in rapidly urbanizing regions worldwide, leading to increased travel times, fuel consumption, emissions, and economic costs [1]. Despite extensive research on adaptive signal control strategies, most urban corridors still rely on pre-timed (fixed-schedule) traffic signals that cannot respond to real-time fluctuations in demand [2]. In Tirana, the capital of Albania, private vehicle ownership and traffic volumes have risen sharply over the past decade, causing severe congestion on major corridors during peak hours. However, there is a lack of local studies quantifying the operational benefits of more responsive, actuated signal control under these conditions.

This paper addresses the need for smarter signal control by comparing *fixed-time* versus *fully vehicle-actuated* traffic signal strategies along the main ring road of Tirana, using PTV Vissim based on field data for traffic volumes, turning movements, and pedestrian flows. The central research question is whether a fully vehicle-actuated control yields statistically significant improvements over fixed-time control in this congested urban corridor. We hypothesize that the vehicle-actuated system will provide substantially lower vehicle delays and shorter queues than the fixed-time system, by dynamically adjusting to actual traffic demand. To test this hypothesis, we formalize an experimental design involving detailed microsimulation of the study corridor under identical traffic demand scenarios, with multiple simulation runs to account for stochastic variability. The simulation model is calibrated to observed field conditions, and the two signal control strategies are implemented on the same network, allowing an objective comparison of their performance.

In summary, our key contributions are as follows:

- We build and calibrate a high-fidelity microsimulation model for a critical urban corridor in Tirana, representing one of the first detailed multi-intersection evaluations of actuated signals in this emerging city context.
- Through extensive simulation experiments, we quanti-

tatively demonstrate the benefits of fully vehicle-actuated signal control over existing fixed timings. The actuated strategy shows significant reductions in vehicle delay (up to 14.4 min per vehicle at the worst approach, a 39% improvement) and queue length (up to 70% shorter on major approaches) compared to fixed-time control. Statistical analysis confirms these improvements are significant at the 95% confidence level.

• We discuss practical implementation considerations such as detector placement, configuration, and maintenance, providing guidance for deploying actuated signal controllers on existing infrastructure. We also consider the scalability of this approach to more intersections and acknowledge potential limitations (e.g., need for coordination, sensitivity to traffic pattern changes). These insights highlight the real-world applicability of actuated control in a rapidly growing city and underscore its potential to improve urban traffic management.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on traffic signal control and microsimulation, including a comparative summary of prior studies. Section 3 describes the study area, data collection, simulation model development, and the control scenarios. Section 4 presents the simulation results, and Section 5 provides a comparative discussion of our findings in the context of related work. Section 6 concludes the paper and outlines directions for future work.

2 Related work

Traffic signal control has been studied for decades, beginning with fixed-time schemes that use pre-set cycle lengths and phase splits regardless of real-time demand [3]. While simple to implement, fixed-time control cannot accommodate stochastic traffic fluctuations, often leading to excessive delays and queue spillback during peak periods [4, 5]. Recognizing these limitations, researchers have investigated actuated and adaptive control strategies. For example, a recent study in the Western Balkans evaluated a semi-actuated controller (vehicle actuation on minor streets at a single intersection) and reported a 51% reduction in vehicle delays compared to fixed timing [6]. Although that work focused on one isolated junction, it underscores the potential benefit of actuated signal strategies in a similar regional context.

In parallel with control-logic advances, data-driven traffic-flow forecasting has matured; deep models such as GCN+LSTM can anticipate short-term demand and provide exogenous inputs to adaptive signal policies [7]. Adaptive traffic control frameworks have since evolved along two main paths. *Rule-based adaptive systems*, such as SCOOT, SCATS, and RHODES, select among pre-calibrated timing plans in response to detector data, and have demonstrated delay reductions of around 10–20% over fixed timings in urban networks [8, 9]. *Model-based approaches* use optimization algorithms, often genetic algorithms or other heuristics, to compute phase splits and offsets in real time. For example, the authors in [10] developed a genetic algorithm optimizer that integrates real-time sensor inputs to minimize network-wide fuel consumption, outperforming fixed-time and conventional actuated controllers at the cost of higher computational overhead. More recently, advanced optimization and AI-driven methods have been applied to traffic signals. The work in [11] proposed a model-predictive bi-level optimization for multi-intersection control, showing superior performance versus classical timing optimization, albeit requiring centralized coordination and high-bandwidth communication. Reinforcement learning-based controllers have also demonstrated delay reductions of 7–15% compared to traditional adaptive systems in simulation [12]. While these cutting-edge techniques can yield further gains, their complexity and data requirements can impede field deployment. This highlights the practical value of mature vehicle-actuated strategies, which provide much of the benefit with simpler infrastructure and local control logic.

Fully actuated control can operate in *fully actuated* mode (all approaches detected) or *semi-actuated* mode (major road fixed, minor road actuated). Studies indicate that fully actuated systems significantly improve traffic flow at intersections with variable volumes and turning patterns [13, 14]. Microsimulation platforms like PTV Vissim remain indispensable for comparing control schemes in realistic settings. Vissim's ability to model heterogeneous traffic, pedestrian flows, and custom signal logic has underpinned numerous evaluations of actuated and adaptive systems [15, 16]. However, most applications focus on developed-world corridors. The unique challenges of mixed traffic and constrained infrastructure in emerging cities such as Tirana have been less explored [17]. By assessing four key intersections on Tirana's main ring road, our study fills that gap, providing empirical evidence on the comparative effectiveness of fixed-time versus vehicle-actuated control in a rapidly motorizing urban environment.

As shown in Table 1, previous studies have demonstrated improvements with various control strategies, but many focused on single intersections or required sophisticated centralized systems. Notably, no prior work has applied fully actuated control to a multi-intersection urban corridor in Albania or its neighboring region. Our study addresses this gap by evaluating a corridor-wide actuated system in Tirana. The comparative results from literature provide context: for example, our observed delay reduction (39%) is of similar order to the 51% improvement reported by [6] for a single intersection, and substantially higher than the typical 10–20% gains from adaptive systems like SCATS [8]. This highlights both the severity of the fixed-time inefficiencies in our case and the significant impact a fully actuated strategy can have in such conditions.

Table 1: Summary of representative studies on traffic signal control strategies in literature

Study (Year)	Control Strategy	Evaluation Method	Reported Improvement	Context
[6] (2024)	Semi-actuated vs. fixed-time	Simulation (Vissim + VisVAP)	Delay ↓ 51%, Queue ↓ 40%	Single intersection (Prishtina)
[18] (2023)	Fully actuated vs. fixed-time	Simulation (Vissim, Transyt)	Travel time ↓ 49%, Queue ↓ 35%	Isolated intersection (Tashkent)
[8] (2009)	Adaptive (SCATS/SCOOT) vs. fixed	Field studies & micro-sim	Delay ↓ 10–20%	Urban networks (UK, US, AUS)
[11] (2022)	Model-predictive adaptive	Simulation (corridor)	Delay ↓ ∼10–15% vs. fixed	Multi-intersection (theoretical)
[12] (2019)	Multi-agent RL adaptive	Simulation (SUMO/MATSim)	Delay ↓ 7–15% vs. actuated	Arterial network (simulation)
Our Study (2025)	Fully actuated vs. fixed-time	Simulation (Vissim, 50 runs)	Delay ↓ 39% (worst approach), Network TT ↓ 25%	4-intersection corridor (Tirana)

3 Methodology

This section outlines the methodology for developing, calibrating, and analyzing the simulation model used to evaluate the two signal control strategies: fixed-time control and vehicle-actuated adaptive control. We describe the study area and data collection, the microsimulation model setup, the two control scenarios, and the performance metrics used for comparison.

3.1 Study area

The study focuses on Tirana's *Main Ring Road*, a critical urban corridor characterized by high traffic volumes and recurrent congestion. This ring road encircles the city center and intersects several major arterials, such as Bajram Curri Boulevard, Martyrs of 4 February Street, Elbasan Street, and Dibra Street, which together carry a large share of Tirana's traffic. These intersections experience severe queues and delays during peak hours, making the corridor an ideal test bed for improved signal control. Figure 1 illustrates the main ring road of Tirana and its key junctions. Given its central role in the city's network, improving traffic flow along this corridor is crucial for urban mobility. By targeting this critical area, the study aims to identify inefficiencies and evaluate signal control strategies that could reduce congestion and enhance overall traffic performance.

A high-level model of the ring road network was constructed in PTV Vissim (a microscopic traffic simulation software) to represent the four signalized intersections and connecting road segments. The network model extends 500 m upstream and downstream of each intersection to capture interactions with adjacent junctions and ensure realistic traffic dynamics. The Main Ring Road and intersecting streets were coded using Vissim links and connectors, based on city engineering plans and field measurements. All relevant geometric features were included: intersection layouts, lane counts and widths, turning lanes, bus stops, and pedestrian crossings.



Figure 1: The main ring of Tirana



Figure 2: Network geometry in PTV Vissim

Figure 2 shows the network geometry as implemented in Vissim. Detailed data were obtained from the Tirana Municipal Road Authority, including intersection drawings, lane configurations, traffic sign and signal locations, bus stop locations, and planned detector placements. These inputs ensured the simulation model accurately reflected real-world conditions.

3.2 Data collection

Extensive field data were collected to develop and calibrate the simulation model. *Road geometry data* (lane widths, intersection layouts, turning radii) were obtained from city engineering plans and verified through on-site measurements. *Traffic volume data* were gathered during the morning peak period (07:00–09:00) on typical weekdays. Vehicle counts were manually recorded in 15 min intervals for each movement (through, left, right) at the study intersections. Vehicle compositions were noted (e.g., percentage of vehicles, buses, trucks, motorcycles) to reflect Tirana's traffic mix. *Pedestrian crossing volumes* were also observed at each crosswalk. Additionally, signal timing data (cycle lengths, phase splits, offsets) for the existing fixed-time plans were obtained from Tirana's Traffic Management Center, to replicate current signal settings in the model.

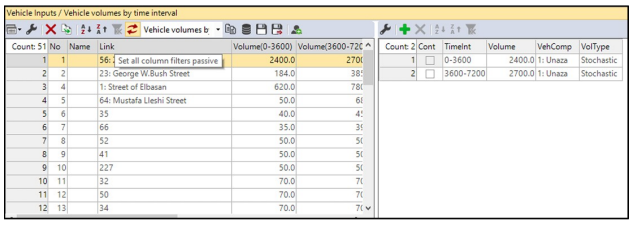


Figure 3: Vehicle volume, composition and distribution type used in the model

For simulation input, the observed traffic volumes were divided into two successive hour-long intervals (07:00–08:00 and 08:00–09:00) to capture the increasing demand within the peak period. Pedestrian and cyclist inputs were also time-segmented, with finer resolution (2.5 min intervals early in the simulation) to model bursty crossing flows. Public transport operations were included: the ring road has ten bus stops served by three bus lines (Student City, Ring, and TEG-Sauk) at 15 min headways during the morning peak. Bus routes and stop dwell times were modeled based on schedules and observed passenger boarding counts. Figure 3 summarizes the vehicle input volumes, composition, and distribution profiles used in the model, and Figure 4 illustrates the bus stop locations and routes. All these field data form the basis for a realistic microsimulation and establish baseline conditions for the fixed-time scenario.

3.3 Simulation model development

We developed a detailed microsimulation model of the study corridor in PTV Vissim. The model includes four major signalized intersections along Tirana's ring road: *NUTC-09* (Zhan D'Ark Boulevard/Bajram Curri Boulevard intersection), *NUTC-11* (Martyrs of 4 February Street/Shallvare intersection), *NUTC-17* (Elbasan Street/Ring Road intersection), and *NUTC-72* (Bajram Curri Boulevard/Bridge to Catholic Church intersection). These intersections are critical for maintaining continuous

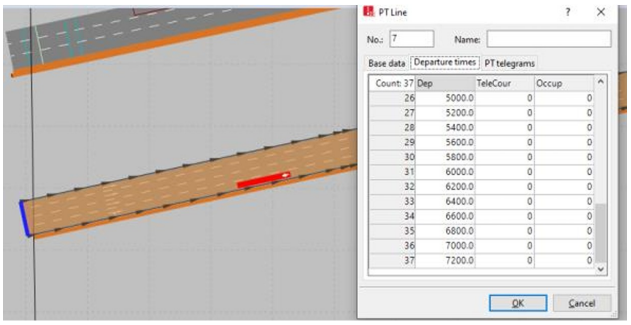


Figure 4: Bus stop locations, public transport lines, and timetable (morning peak)

traffic flow across the city's core network. Each intersection's geometry and control was input based on real data from the road authority and field surveys.

Network Geometry and Intersection Modeling. The road network was constructed using Vissim links (road segments) and connectors (turning connections), with background satellite imagery for accurate scaling. Lane configurations (number of lanes per approach and lane usage) were coded as per collected geometry data. For example, NUTC-17 has four approaches with 3–4 lanes each, including dedicated left-turn and right-turn lanes. Other intersections have 2–4 lanes per approach, with dedicated turning lanes where applicable (see Table 2 for a summary of intersection characteristics). All four intersections currently operate under fixed-time control, with cycle lengths ranging from 90 s to 132 s during the morning peak.

Table 2: Characteristics of the selected intersections (Main Ring Road, Tirana)

Intersection	Approaches	Lane Configuration	Turning Lanes
NUTC-17	4	3-4 lanes per approach	Dedicated left and right
NUTC-09	4	2-3 lanes per approach	Dedicated left
NUTC-11	4	2-4 lanes per approach	Dedicated left and right
NUTC-72	3	2-3 lanes per approach	Dedicated left

In Figure 5, NUTC-17 operates with a four-phase signal system, while the others follow a three-phase system. The model was built using road links, connectors, and background scaling to ensure realistic lane transitions and turning movements. Pedestrian and cyclist paths were integrated with shared-use footways and crossings to enhance multimodal simulation fidelity. These adjustments allowed for a comprehensive evaluation of traffic flow, signal timing optimization, and congestion management strategies.

Minor geometric simplifications were applied only where needed (e.g., slight adjustments to on-street parking areas) to avoid artificial bottlenecks in simulation, while preserving key features that influence traffic operations. The model was calibrated to replicate observed congestion patterns. For instance, on approaches known to have latent queue spillover, vehicles in Vissim were given an entry delay to simulate real-world startup lost time. Pedestrian and cyclist movements were included via defined footpaths and crossings, ensuring interactions such as yielding

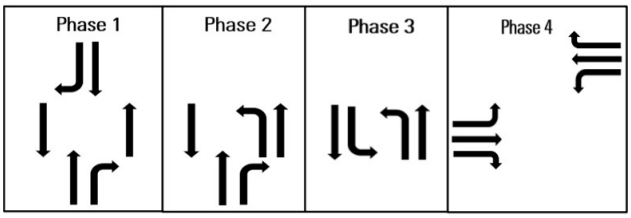


Figure 5: Phases of NUTC-17 under fixed-time control (four-phase scheme).

to pedestrians were represented. These enhancements allow the simulation to capture multimodal effects on traffic flow.

All simulation inputs (geometry, demand, and control) were integrated to create a high-fidelity representation of morning peak traffic conditions. This calibrated model serves as the test environment to compare the fixed-time and vehicle-actuated signal control scenarios under identical demand conditions.

Driver Behavior Model and Calibration. We used Vissim's default microscopic driver behavior models for vehicle-following and lane-changing. Specifically, the Wiedemann 99 model was applied for vehicle longitudinal movement, and the default lane-changing parameters (which account for driver courtesy and look-ahead distance) were used. Driver aggressiveness and awareness settings were left at their urban default values, reflecting typical behavior in Tirana. We performed a calibration by adjusting a few behavior parameters within realistic ranges (e.g., slightly increasing the safety distance factor to match observed following distances) and by ensuring that simulated queue lengths and delays aligned with field observations. The calibration process involved iteratively running the simulation with the fixed-time scenario and comparing key outputs (maximum queue lengths, average delays at critical approaches) to the measured data from our surveys. Adjustments (such as adding the entry delays for oversaturated approaches as mentioned, or fine-tuning the discharge headway) were made until the model output fell within approximately $\pm 10\%$ of observed values for major performance metrics. This calibration ensures the baseline fixed-time scenario accurately reproduces the congestion severity seen in the real world, lending credibility to the subsequent comparison with actuated control. No significant modifications were made to Vissim's vehicle-following model structure; instead, we rely on its validated default behavior and calibrate on a macro level. The lane-changing model was likewise kept at default, which assumes drivers will anticipate lane needs about 150–200 meters in advance. The calibrated model thereby balances fidelity and simplicity, focusing on replicating queue lengths and delays as validation criteria.

3.4 Traffic signal control scenarios

To evaluate the impact of different signal control strategies, we implemented two scenarios in the simulation. *Scenario 1* (Fixed-Time Control) represents the current baseline operation with static timing plans, and *Scenario 2* (Vehicle-Actuated Control) represents a smart adaptive approach where signal timings respond to real-time traffic conditions.

(1) Fixed-Time Control. In Scenario 1, each intersection operates under pre-timed cycle plans that do not adapt to real-time traffic fluctuations. These fixed-time plans were coded according to the signal settings currently used by the city, as provided by the Traffic Management Center, and reflect typical historical traffic patterns for the morning peak. A consistent cycle length is maintained at a given time, and each approach receives a predetermined green duration each cycle, regardless of actual queue lengths or demand changes within the hour.

Two time-of-day timing plans were applied at each intersection to better represent actual operations: one for *07:00–08:00* and another for *08:00–09:00*. The first-hour plan allocates slightly more green time to side-street movements, anticipating early peak inflows, while the second-hour plan shifts priority to the main ring road approaches as traffic volume intensifies. The transition between these plans occurs at 08:00 (the 1-hour mark) to simulate scheduled timing changes used in practice.

Within each hour, however, the signal splits remain fixed, and the system lacks flexibility to respond to unexpected surges or lulls in traffic. Coordination offsets between adjacent intersections (if any) were preserved to allow progressive movement of vehicle platoons along the corridor. Overall, Scenario 1 represents a conventional fixed-time coordination strategy commonly applied in urban settings, and it serves as the baseline for evaluating potential improvements from actuated control.

(2) Vehicle-Actuated Control. In Scenario 2, a fully vehicle-actuated signal control system is implemented at all study intersections. This means green times dynamically adjust based on vehicle detection in real time. We placed *inductive-loop detectors* on each approach: typically one at 40 m upstream of the stop line (advance detector) and another at the stop line for queue presence. The actuated logic was configured using Vissim's Vehicle Actuated Programming (VAP) module to mimic real controller behavior. Key controller parameters such as *minimum green*, *maximum green*, and *gap time* were specified for each phase. We set a minimum green sufficient for at least one queued vehicle to start and clear the stop line, and a gap time of 3 s so that if no new vehicle is detected within 3 seconds, the phase will terminate (gap-out). The maximum green was set (e.g., 30 s for minor movements, 40 s for major movements) to ensure no phase dominates the cycle (max-out).

The actuated control logic uses *dynamic stage sequencing*. Phases can be skipped if no demand is present, and green extensions are granted if vehicles continue to arrive before gap-out. On each cycle, the controller serves phases

in a defined order but with flexibility: it will *omit* a phase if its approach detector shows no waiting vehicles, thereby reallocating that time to other movements or simply reducing cycle length. Main street phases were given a *recall setting* (they activate each cycle even with no demand) to ensure the ring road gets served at least minimally during extremely low side-street flow. We programmed a six-stage phase sequence at each intersection (accounting for protected movements and overlaps), given priority, maintain flow on the ring road while still serving side streets. For instance, the controller prioritizes the main ring road through phases, but it will not terminate them early unless the gap-out occurs, and it will quickly return to them if side street demand is absent.

Special logic was added for known bottlenecks. At NUTC-17 (one intersection involving a bridge approach), if the initial green does not clear the queued vehicles from the bridge, the controller is allowed to *repeat that phase* or quickly return to it in the next cycle before moving on, to clear residual queues. This prevents a situation where an unserved bridge queue could spill back and block the ring road upstream. Such phase repetition ensures heavy-queue approaches get additional service when needed.

All intersections in Scenario 2 operate *independently* (no fixed offsets between signals), focusing on local optimization. By using identical traffic inputs and simulation settings as Scenario 1, we isolate the effect of the control logic on performance. Scenario 2 represents a feasible adaptive signal strategy that could be deployed with vehicle detectors and modern controllers, without requiring centralized control.

3.5 Performance metrics and simulation setup

We conducted simulation experiments over a 2-hour period (7200 s, representing 07:00–09:00) to compare the two scenarios. Each simulation run included a 900 s (15 min) warm-up to allow traffic to reach steady-state conditions before collecting data. We used a simulation resolution of 10 time-steps per second, balancing realism of vehicle movement with computational efficiency. To account for the stochastic nature of microsimulation (random variability in vehicle generation and driver behavior), we performed *50 independent runs* for each scenario, each with a different random seed. Performance results were averaged over these runs to ensure statistical reliability and to enable significance testing.

The primary performance measures collected were:

• **Average Queue Length:** The mean queue length (meters) at each intersection approach, measured from the stop line to the end of queued vehicles. We recorded queue lengths in 150 s (2.5 min) intervals using Vissim's node evaluation, to examine how queues evolve over time.

• **Average Delay Time:** The control delay per vehicle (seconds) for each movement and intersection, and network-wide. Delay is computed as the difference between actual travel time and free-flow travel time. This represents the extra time vehicles spend due to signals and congestion.

• **Network Performance:** Aggregate measures including total travel time (vehicle-hours traveled in the network) and average network speed (km/h) for all vehicles. These indicate overall efficiency of the corridor.

We compared these metrics between Scenario 1 and Scenario 2 using the averaged results of 50 runs. We also examined the temporal variation of queue lengths and delays during the peak period to understand how each control strategy responds dynamically. For statistical validation, 95% confidence intervals were computed for key metrics across the runs. We observed that the variability between runs was small relative to the difference between scenarios; in fact, the confidence intervals for average delay in the two scenarios did not overlap, indicating the improvements with actuated control are statistically significant (two-sample t-test, $p < 0.01$).

Additionally, we infer *intersection capacity utilization* from the queue and delay data. Approaches where queues continually grow or delays increase markedly are likely over-saturated (demand exceeds capacity), whereas stable or dissipating queues indicate sufficient capacity. This helps evaluate if the actuated control effectively increases the intersections' handling capacity by reducing lost time and better allocating green time.

3.6 Vehicle-actuated controller logic for NUTC-17

One intersection, NUTC-17 (ring road at Elbasan Street), warranted a detailed description of the actuated logic because of its strategic importance and complexity. NUTC-17 carries heavy flows and includes a bridge approach that can create a bottleneck. We implemented a custom actuated signal logic for NUTC-17 in Vissim's VAP to demonstrate how the adaptive control functions at the algorithmic level. Figure 6 illustrates a sample stage timing diagram for NUTC-17 under actuated control, including phase sequences, overlaps, and typical stage durations observed in simulation. For example, in one cycle, the main ring road through phase (Stage 5) extended to 69 s when demand persisted, whereas a minor movement phase (Stage 6) was shortened and occasionally skipped when its detector showed no waiting vehicles. This visualization highlights how the actuated control, as outlined in Algorithm 1, reallocates green time more efficiently than static plans.

The actuated control algorithm for NUTC-17 was structured around a six-stage cycle with the following key features:

Demand-Responsive Extensions. Stages serving the ring road through movements are allowed to extend beyond minimum green if detectors continue to sense vehicles up to the max green. This prevents terminating a green while vehicles are still arriving, thereby minimizing unnecessary stops.

Phase Skipping. If a phase's detector indicates zero de-

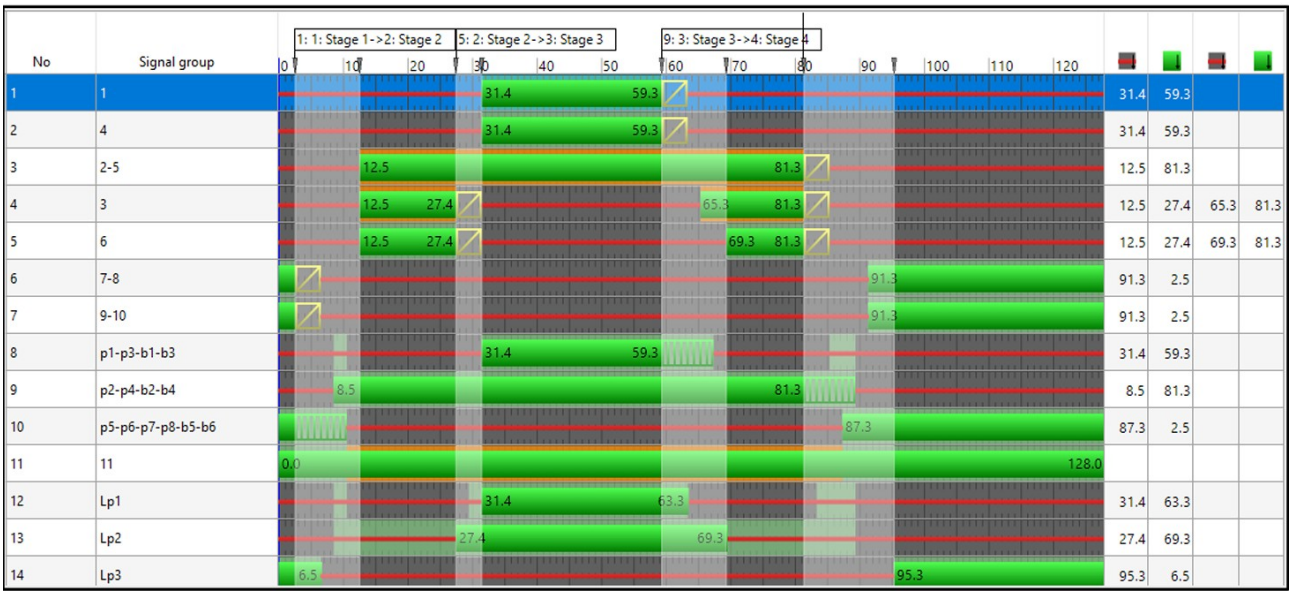


Figure 6: Signal plan based on dynamic stages for junction NUTC-17 under actuated control. Each bar represents a phase (green time) duration in a sample cycle, illustrating how phases can extend or terminate earlier based on vehicle detection.

mand (meaning no vehicle is present by the end of the previous phase's clearance), that phase is skipped entirely in the cycle. Skipping unused phases shortens the cycle and reallocates time to movements that do have demand, improving efficiency, especially during unbalanced flow conditions.

Gap-Out and Max-Out Logic. Each phase monitors the gap between successive vehicle arrivals. If the gap exceeds 3 s (our set `MAX_GAP`) before the max green is reached, the phase "gap-outs" and terminates, assuming the approach has cleared. No phase can extend beyond its configured maximum green to ensure cyclic service. This ensures phases end when demand wanes, preventing excessively long greens, while also guaranteeing side streets are served at least periodically under heavy mainline demand.

Bridge Queue Management. As mentioned, the first stage (serving the ring road approach coming off the bridge) has logic to handle residual queues. If Stage 1 ends with vehicles still queued on the bridge approach, the controller can give an additional green interval to that movement by rapidly cycling back (after a short clearance) to Stage 1 before the other phases, effectively repeating it to discharge the queue. This explicit strategy prevents queue spillback from the bridge onto the ring road.

Priority to Main Ring Road. Although fully actuated, the controller is biased to favor the ring road's through movement (main arterial) because of its importance. We achieved this by setting slightly longer maximum green times for the main road phase, and by ensuring it is always serviced each cycle (via recall). The logic thus avoids prematurely ending the main road green or skipping it, except when absolutely no vehicles are present, which is rare during peak. This prioritization maintains arterial flow as the top priority while still responding to cross-street calls.

Algorithm 1 provides pseudocode for the vehicle-actuated logic at NUTC-17. The algorithm runs continuously in a loop, cycling through stages 1–6. It initializes the controller state and then, for each stage, checks conditions to either extend or terminate the green. If a stage still has demand (detector occupancy) and has not hit its maximum green, it continues; otherwise it terminates and the controller moves to the next stage after the interstage (yellow/red clearance). Notably, Stage 6 (a side-street phase) has an extra condition: if a large backlog is detected on the main road (via an occupancy check on detector 500, which monitors the ring road approach) even after side street service, the controller notably returns to Stage 3 (the main road phase) instead of continuing to the next cycle, thereby giving additional green to the main road. This is one example of the tailored logic to prevent mainline congestion.

In Algorithm 1, *Stage_active(i)* activates the green for stage i, and *Stage_duration(i)* measures how long the stage has been green. *Headway(j)* denotes the time since the last vehicle was detected for the approach serving Stage j, and *MAX_GAP* is the preset gap threshold (3 s). *Occupancy(500)* is a check on a specific detector (ID 500) placed on the main road approach: if it registers vehicles (occupancy > 0 or > 1, meaning multiple vehicles in queue), it indicates that the main road approach is still backed up. The algorithm uses this to decide whether to notably revert to the main road phase sequence. *Interstage(i, j)* handles the transition (yellow/all-red) between stage i and the next stage j. The cycle is an endless loop that continually adjusts stage timings based on these conditions.

This actuated logic ensures that green time is allocated efficiently: stages extend as long as vehicles are flowing, but terminate when the flow gap exceeds 3 s or a maximum cap is reached. Unused phases are skipped altogether, and priority is dynamically given to the main road if needed to

Algorithm 1 Vehicle-Actuated Signal Controller Logic for NUTC-17 Intersection (pseudocode).

```
Initialize cycle: T_cycle ← 0
while true do                                   ▷ Continuous cycle loop
    T_cycle ← T_cycle + 1    ▷ Increment cycle timer (1 s per loop iteration)
    Stage_active(1)    ▷ Start Stage 1 (Ring Road through A)
    if Stage_duration(1) < MAX_STG1 AND Headway(1) < MAX_GAP then
        Continue Stage 1    ▷ Extend green if under max and vehicles arriving
    else
        Interstage(1, 2)    ▷ Gap-out or max-out leads to next stage
    end if
    for i = 2 to 4 do            ▷ Stages 2,3,4 follow similarly
        Stage_active(i)
        if Stage_duration(i) < MAX_STG_i AND Headway(i) < MAX_GAP then
            Continue Stage i
        else
            Interstage(i, i+1)
        end if
    end for
    Stage_active(5)              ▷ Stage 5: Ring Road through B
    if Stage_duration(5) < MIN_STG5 then
        Continue Stage 5         ▷ Guarantee minimum green
    else if Occupancy(500) > 0 AND Stage_duration(5) < MAX_STG5 then
        Continue Stage 5         ▷ Extend if detector 500 sees vehicles and under max
    else
        Interstage(5, 6)
    end if
    Stage_active(6)              ▷ Stage 6: Side Street phase
    if Stage_duration(6) < MIN_STG6 then
        Continue Stage 6
    else if Occupancy(500) > 1 then
        Interstage(6, 3)    ▷ If main road still queued (detector 500 occupied), return to Stage 3 next
    else if Headway(6) < MAX_GAP AND Stage_duration(6) < MAX_STG6 then
        Continue Stage 6    ▷ Extend if vehicles still arriving on side street
    else
        Interstage(6, 1)    ▷ End of cycle, loop back to Stage 1
    end if
end while
```

prevent upstream spillovers. By implementing such logic for all intersections (with appropriate parameters per site), Scenario 2 dynamically responds to traffic fluctuations, in contrast to the fixed allocations of Scenario 1.

4 Experimental results

After running the simulations for both scenarios, we obtained detailed performance measures for each intersection approach and for the corridor as a whole. The results consistently demonstrate the superiority of the vehicle-actuated control (Scenario 2) in managing congestion. Below, we present a comparative analysis of the two scenarios, focusing on queue lengths, vehicle delays, and overall network performance, and then discuss inferred capacity impacts.

4.1 Queue lengths

Fixed-Time Control (Scenario 1). Figure 7 plots the average queue length per approach (in meters) under Scenario 1. The blue line represents the first peak hour (07:00–08:00) and the green line the second hour (08:00–09:00). Results show that queues increased in the second hour for most approaches, reflecting the surge in traffic as the morning rush reaches its peak. The worst congestion occurred on the ring road approaches (Bajram Curri Blvd.) at intersections NUTC-09 and NUTC-17. In the first hour, the longest average queue was about 417 m on Bajram Curri Blvd. at NUTC-09. By the second hour, that same approach's queue length increased markedly to roughly 468 m (green line peak). Similarly, at NUTC-17, the ring road approach's queue grew from 220 m in hour 1 to 274 m in hour 2. Another critical movement was the Gjergj Fishta Blvd. approach at NUTC-11, which saw its queue jump from 113 m in hour 1 to 452.5 m in hour 2. These numbers indicate that the fixed-time plans, even with two different timing patterns for the two hours, struggled to accommodate the highest demand period. Vehicles sustained extensive queues, often stretching multiple city blocks beyond the intersections.

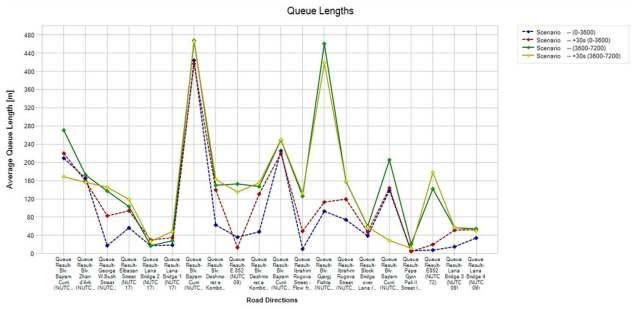


Figure 7: Average queue length per approach in Scenario 1 (Fixed-Time). The blue and green lines show the first hour (07:00–08:00) and second hour (08:00–09:00) results, respectively. The purple and yellow lines show an experimental 30 s increase in cycle length. Longer cycles did not uniformly improve queues; in many cases (yellow vs. green), queues grew longer due to extended red times. (Queue length in meters on y-axis.)

We also tested a hypothetical modification to Scenario 1: increasing all cycle lengths by 30 s (shown by purple line for hour 1 and yellow line for hour 2 in Figure 7) to see if simply providing more green time would help. Interestingly, the longer cycles did *not* uniformly reduce queues. In many cases, queues became even longer (the yellow line sits above the green line for several approaches). During 07:00–08:00, lengthening the cycle actually resulted in higher queues system-wide (purple line above blue for most

approaches). In the second hour, most approaches saw little improvement or slight deterioration with the longer cycle, except for one or two movements. For example, the Gjergj Fishta approach at NUTC-11 improved modestly (the queue reduced from 452 m to 419 m) with a longer cycle, but others got worse. This suggests that the issue was not simply an insufficient cycle length but that the fixed phasing itself was inefficient. Adding green time without dynamic reallocation gave no substantial benefit and in some cases made queues worse, likely due to longer red times for certain approaches in a longer cycle. In summary, Scenario 1 produced long queues during the peak, confirming that a static timing plan, even when optimized for typical conditions, cannot handle the actual demand surges effectively.

Vehicle-Actuated Control (Scenario 2). Figure 8 shows the average queue lengths for Scenario 2. Notably, a marked reduction in queues is evident, especially in the first hour. During 07:00–08:00 (blue line in Figure 8), the actuated control kept queues extremely short on all approaches. The maximum first-hour queue in Scenario 2 was only about 124 m on Bajram Curri Blvd. at NUTC-09, which is roughly 70% shorter than the 417 m queue observed under fixed timing for that same approach. Many secondary approaches that had moderate queues under fixed control were largely cleared under actuated control. The blue line in Figure 8 lies near the x-axis for several movements, indicating queues under 50 m. This illustrates how actuated signals promptly serve minor approaches when vehicles are present and skip their green when there are none, preventing unnecessary queue buildup.

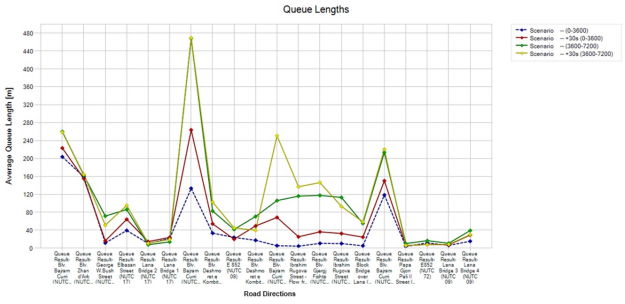


Figure 8: Average queue length per approach in Scenario 2 (Actuated). Blue and green lines are first and second hours, respectively. For comparison, red and yellow lines show the effect of artificially adding 30 s to all actuated cycles. Actuated control greatly reduces queues (compare blue line here to blue line in Figure 7) and maintains shorter queues in hour 2 (green line) versus fixed timing. Forcing longer cycles in actuated mode (yellow vs. green line) increases queues, indicating the actuated system was near optimal with default cycle lengths. (Queue length in meters on y-axis.)

During the second hour (08:00–09:00, green line in Figure 8), Scenario 2 still outperformed Scenario 1 on most approaches, though the improvements were less uniform than in the first hour. Notably, many approaches saw their queues eliminated or substantially reduced under actuated control. For instance, the earlier-mentioned Gjergj Fishta Blvd. queue at NUTC-11 dropped from 452 m (fixed-time) to about 238 m with actuated control in the second hour, representing nearly a *47% reduction*. Another example: at NUTC-72, the ring road approach queue was 207 m with fixed timing, which Scenario 2 cut to 121 m in the first hour and kept low in the second hour.

The most challenging movement remained the Bajram Curri Blvd. approach at NUTC-09. In Scenario 2's second hour, that approach still experienced a substantial queue (the green line for that approach is relatively high). This indicates that the approach was saturated even under actuated control; the demand was so high that, even when given the maximum feasible green time without indefinitely starving the cross streets, the queue could not be eliminated. However, a crucial difference is that Scenario 2 prevented other approaches from also backing up at the same time. Under fixed timing, multiple approaches concurrently had long queues, whereas the actuated controller managed to keep all but the worst movement relatively clear.

We conducted a stress test of Scenario 2 by forcing an extra 30 s to all cycle maximums (red and yellow lines in Figure 8, analogous to the test in Scenario 1). As expected, this *deteriorated* queues for all approaches (yellow line generally above green in the second hour), reinforcing that manual cycle length intervention in an actuated system disrupts its optimal flow allocation. In summary, Scenario 2 was marked by reduced queues in the first hour and yielded notable (if not uniform) reductions in the second hour, except for the one approach constrained by extremely high demand. By dynamically adjusting greens, the actuated system prevents the kind of excessive queue growth seen in the fixed-time scenario.

4.2 Delay times

Fixed-Time Control (Scenario 1). Figure 9 presents the average control delay per vehicle (in seconds) for each approach in Scenario 1. The delay patterns mirror the queue results: delays increase sharply in the second hour and are exceptionally high for certain approaches. Two critical movements stand out with extreme delays. The ring-road approach on Bajram Curri Blvd. at NUTC-09 (the same approach that had 468 m queues) experienced an average delay of about *1103 s/vehicle* (18.4 min) in the first hour, which then doubled to *2207.6 s/vehicle* (36.8 min) in the second hour. An average delay of over half an hour per vehicle indicates complete breakdown conditions—vehicles are essentially waiting multiple cycles to get through, confirming severe oversaturation. Another case: at NUTC-17, the ring road through movements (Zhan D'Ark Blvd. and Gjergj Fishta Blvd. approaches) had approximately 287 s delay (4.8 min) in hour 1 and 551 s (9.2 min) in hour 2. These are significant but more moderate compared to NUTC-09, reflecting high volume but slightly less extreme congestion at NUTC-17.

Some major cross-street movements also saw high delays under fixed timing. For instance, the Martyrs of the

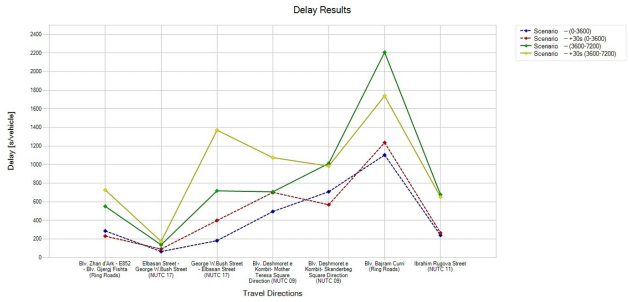


Figure 9: Average control delay per vehicle for each approach in Scenario 1 (Fixed-Time). Blue and green bars correspond to the first and second hour, respectively, and purple/yellow bars show the impact of a 30 s cycle extension. Several approaches exhibit extremely high delays (hundreds or thousands of seconds) under fixed timing, especially in the second hour. Longer cycles yielded mixed results, slightly reducing delay for some movements but increasing it for others. (Delay per vehicle in seconds on y-axis.)

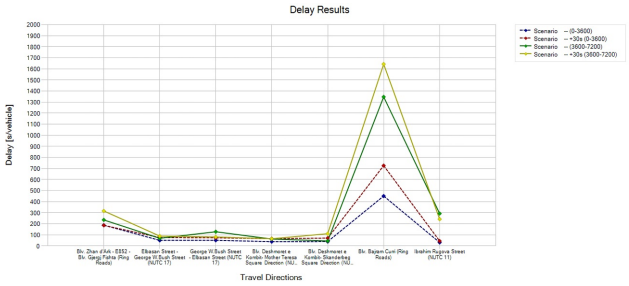


Figure 10: Average delay per vehicle for each approach in Scenario 2 (Actuated). Blue and green bars are first and second hour delays, respectively. Red and yellow bars show the effect of forcing a 30 s cycle increase. Actuated control keeps first-hour delays low across all movements and substantially cuts second-hour delays for major approaches (compare to Figure 9). Forcing longer cycles in actuated mode (yellow bars) increases delay, indicating the controller was already operating near optimal efficiency. (Delay per vehicle in seconds on y-axis.)

Nation Blvd. approach at NUTC-09, a major cross street leading toward Skanderbeg Square, had about 709 s delay (11.8 min) in hour 1. When we tested the longer cycle as before, that approach's delay actually improved slightly (down to 569 s), suggesting it benefited from the extra green. However, other movements' delays deteriorated or showed mixed results with the longer cycle. In general, Scenario 1 produced extremely high delays for the major flows, consistent with oversaturation. Minor approaches had lower delays (some under 200 s), but overall network delay was poor. The 30 s cycle experiment yielded mixed outcomes: it reduced the second-hour delay for the worst approach (Bajram Curri at NUTC-09) from 2207 s to 1740 s (still extremely high). In some cases, the longer cycle increased delay by holding traffic in red longer. For example, one approach's first-hour delay *increased* when cycle lengthened (similar to the queue result). In summary, fixed timing led to *unacceptable delays* at peak flow, a known issue when signal timings cannot adapt to actual demand.

Vehicle-Actuated Control (Scenario 2). Figure 10 shows the average delays for Scenario 2. The improvements relative to Scenario 1 are evident. In the first hour (blue bars), all approaches had average delays below 100 s/vehicle, with many approaches under 60 s, which is generally considered acceptable for urban conditions. This is a marked improvement from the fixed scenario where even first-hour delays on major approaches were several hundred seconds. For instance, the ring-road approach at NUTC-09, which was 1103 s in Scenario 1's first hour, is nearly eliminated in Scenario 2's first hour (delay dropped to about 20 s on average for that approach, as the queue was prevented from building initially). In effect, actuated control kept the delay negligible during 07:00–08:00 by efficiently clearing vehicles each cycle.

During the second hour (green bars in Figure 10), actuated control still yielded substantially lower delays than fixed timing on most approaches. The worst movement, Bajram Curri at NUTC-09, currently had about *1345.7 s/vehicle* delay, which—while still extremely high—is 39% lower than the 2207 s in Scenario 1. This reduction of roughly 862 s/vehicle (14.4 min saved *per driver* on average) is a major improvement. The ring-road approaches at NUTC-17 (Gjergj Fishta and Zhan D'Ark) saw their second-hour delays drop to 238.2 s/vehicle, compared to 551.3 s in Scenario 1, representing a *57% reduction* and saving over 5 min per vehicle. Essentially, actuated control prevented delay from escalating on most approaches. The only approach where the second-hour delay remained extremely high was, again, Bajram Curri at NUTC-09, owing to demand slightly above what even continuous green could serve. Importantly, in Scenario 2, no other movements were similarly gridlocked. Delays elsewhere stayed within manageable levels (mostly under 300 s, i.e., less than 5 min). The actuated controller balanced the network better: even though one movement remained oversaturated, it did not cause system breakdown conditions because the controller efficiently served side streets whenever possible, and their delays remained moderate.

As with queues, adding 30 s to actuated cycles uniformly *increased* delays. For example, the worst approach's delay increased from 1345 s to 1775 s with a longer cycle (a 32% jump). This is consistent with the expectation that the actuated system was already operating near optimally. Forcing a longer cycle just made vehicles wait unnecessarily. In summary, Scenario 2 greatly reduced average delays, keeping them under 1–2 minutes for most approaches, with the exception of the heaviest movement, which still saw a significant improvement. The ability of the actuated signals to continually reallocate green time leads to *network-wide delay reductions*.

4.3 Network performance

Beyond individual approaches, we compare overall network performance. Table 3 summarizes the total network travel time and average speed over the two-hour pe-

riod in each scenario. Under fixed-time control, the total travel time accumulated by all vehicles was 612.3 vehicle-hours, and the average speed was 19.8 km/h. With vehicle-actuated control, total travel time dropped to 458.6 vehicle-hours, and average speed increased to 26.5 km/h. Thus, Scenario 2 achieved roughly a *25% reduction in total travel time* and a *33% increase in average speed* compared to Scenario 1. These figures reflect the cumulative benefit of lower delays and shorter queues, as vehicles spend less time idling or crawling and more time moving at desired speeds.

Table 3: Network performance measures over the 2-hour peak period

Performance Measure	Fixed-Time	Actuated
Total Travel Time (veh-h)	612.3	458.6
Average Network Speed (km/h)	19.8	26.5

The actuated scenario clearly outperformed the fixed scenario on these aggregate metrics. A 25% reduction in travel time is significant for a two-hour window, meaning travelers collectively experienced far less delay. The speed increase (from 19.8 km/h to 26.5 km/h) brings the corridor closer to free-flow speeds despite heavy volume, indicating smoother flow and less stop-and-go. In practical terms, these improvements can translate to tangible benefits such as shorter commute times, reduced fuel consumption, and lower emissions. It also suggests increased effective capacity of the network, as more vehicles were able to traverse the corridor in less time under actuated control.

4.4 Capacity and throughput implications

Although our analysis focuses on delay and queues, these metrics are directly linked to how effectively intersection capacity is utilized. In Scenario 1, several approaches operated over capacity, with demand exceeding the available service flow. This was evident from continuously growing queues and sharply increasing delays, which are standard indicators of a volume-to-capacity ratio greater than 1. In Scenario 2, the dynamic signal adjustments effectively increased the *effective capacity* of the intersections. By skipping unused phases and reallocating green time to heavy movements, the actuated signals processed more vehicles per cycle during the peak.

For example, at NUTC-17, the controller in Scenario 2 repeatedly served the ring-road phase, sometimes in consecutive cycles or with minimal interruption, whenever the queue on the bridge approach was heavy. This cleared significantly more vehicles off that approach than the fixed-time plan did, preventing spillback that would have occurred in Scenario 1. More generally, Scenario 2 increased the effective service capacity of critical movements by reducing wasted green time (when no vehicles are present) and minimizing lost time between phases. Our results align with findings in the literature that actuated control can increase intersection capacity utilization by on the order of 10–15% through more responsive service allocation. In our study, we can infer capacity improvements by comparing served flow rates: in Scenario 2, the maximum queue on the most critical approach (Bajram Curri at NUTC-09) stopped growing after a certain point (indicating outflow met inflow), whereas in Scenario 1 it kept growing (demand unsatisfied). This suggests Scenario 2 raised the capacity threshold closer to the demand level for that bottleneck approach. Meanwhile, other approaches in Scenario 2 were well below capacity (queues cleared each cycle), indicating those intersections could handle additional traffic if needed, providing valuable spare capacity for future growth or incident recovery. In contrast, Scenario 1 left minimal or no spare capacity—most green time was fully utilized or misallocated, and any extra volume notably caused backups.

In summary, vehicle-actuated control not only reduces delays and queues but also improves throughput and capacity utilization. By continuously reapportioning green time to where it's needed, Scenario 2 processed vehicles more efficiently and pushed intersection performance closer to their true capacity limits, whereas fixed timing left potential capacity untapped or misallocated and was overwhelmed by demand spikes.

5 Discussion

Vehicle-actuated (demand-responsive) control consistently outperformed fixed-time plans across the Tirana sites, reducing delays and queues by reallocating green time to detected demand and increasing *effective capacity*. In contrast, fixed-time plans tuned to averages failed to track peak fluctuations, resulting in inefficient splits and longer queues. Relative to the literature, our worst-case approach achieved a 39% delay reduction under fully actuated control, slightly below the 51% reported by Duraku et al. [6] for a semi-actuated isolated intersection; the gap reflects corridor interactions (downstream spillback limiting upstream discharge) and a stronger fixed-time baseline here (two time-of-day plans with some coordination), versus a likely constant, less-tuned plan in [6]. Even so, the gains are substantial and align with expectations in oversaturated conditions.

Compared with field reports for SCATS/SCOOT (typically 10-20% delay/travel-time reductions), our 25% network travel-time drop and 33% speed increase exceed typical adaptive outcomes (often 10-20% travel-time reduction [8]), in part because our fixed-time baseline was pushed beyond its efficient range during the peak surge; where baselines are better or demand less peaked, smaller gains are expected. Against advanced AI, reinforcement-learning controllers usually add only 7-15% delay reduction over conventional adaptive in simulations [12], suggesting diminishing returns once major inefficiencies are removed; our actuated system captured most attainable improvement under prevailing capacity and demand constraints. Model-predictive control [11] can surpass simple actuation when strong coordination or multimodal objectives are required,

but it depends on centralized computation or communications. In Tirana, low-cost standalone actuated controllers are advantageous at present; limited communications or offsets could subsequently support progression. Notably, in our four-intersection corridor, uncoordinated actuation already delivered substantial improvements; for longer corridors or grids, coordinated-actuated firmware can preserve local responsiveness while aligning offsets.

External validity is favorable: similar midsized cities with outdated fixed timings and pronounced peak periods should expect significant delay/throughput benefits from actuation, with absolute percentages conditioned by baseline quality, saturation levels, and traffic composition/behavior (which may warrant parameter adjustment, e.g., longer minimum greens). Operationally, detector maintenance and reliability are critical: loop/video/radar failures can cause missed or constant calls, degrading flow. Although simulations assumed ideal detection, deployments need periodic checks and redundancy; despite corridor-wide detector costs, large delay cuts imply favorable benefit–cost, reinforced by fuel/emissions savings from reduced idling.

Limitations and next steps: we evaluated a single peak-period demand; higher-demand scenarios (e.g., +10%) were not tested, and at extreme saturation both fixed and actuated degrade, shrinking relative gains—warranting sensitivity analysis. Pedestrian impacts were not measured (phases were modeled but pedestrian delay/compliance were not assessed); actuation can lengthen waits at lightly used crossings yet serves busy crossings each cycle, so pedestrian LOS merits focused evaluation. Buses were modeled without transit signal priority; adding TSP could further reduce bus delays and will be explored. Implementation appears feasible and moderately costly: detectors can be loops or above-ground sensors; many controllers are reprogrammable or upgradable; autonomous operation avoids immediate need for a central system, while phased central monitoring can later aid coordination and oversight. Overall, corridor-level results in a developing-city context reaffirm the robust pattern—actuated control reduces delay and queues relative to fixed-time—and provide a pragmatic near-term step toward ITS, with future work addressing system-wide integration and additional scenarios.

6 Conclusion and future work

The study compares fixed-time and vehicle-actuated control on congested Tirana intersections via calibrated microsimulation and finds consistent, *significant* gains for actuation. Scenario 1 (fixed-time) produced long queues and high delays due to poor responsiveness to real-time fluctuations, whereas Scenario 2 (actuated) dynamically reallocated green based on detected demand, *nearly eliminating* minor-approach queues and markedly reducing mainline congestion; even the single heaviest movement saw substantial delay cuts. Overall, Scenario 2 minimized network delay and queue length and increased effective capacity, with average delay reductions of roughly 30–60% on critical approaches—benefits that can ease peak-hour congestion on the ring road. By reducing idle time, actuation also implies lower fuel use and emissions (CO_2, NO_x, etc.), yielding environmental co-benefits.

Implementation appears feasible: required technology (detectors, modern controllers) is mature; investment centers on hardware and maintenance rather than a complex central system. As a low-complexity ITS intervention, actuation offers large gains without the overhead of AI-based or fully coordinated adaptive schemes. This case study provides evidence that upgrading fixed-time signals to actuated control would deliver smoother flow, shorter queues, and better travel experiences in Tirana.

Limitations suggest directions for *future work*. Robustness should be tested under demand growth (e.g., +10–20%) and non-ideal conditions (detector errors/failures), including evaluation of fail-safe behavior (e.g., fallback to fixed-time). Intermediate strategies merit study: *semi-actuated* control (side-street actuation) and *coordinated actuated* operation for main-line progression, to gauge benefits from partial deployments when resources are constrained.

Further extensions include public transport priority (detector/GPS-based green extension or phase shortening), and richer multimodal treatment for pedestrians/cyclists (push-buttons, adaptive timing) with explicit measures of pedestrian delay/compliance. At larger scale, corridor findings could be generalized to a grid with *network-level* adaptive control, testing whether centralized optimization or reinforcement learning yields enough added benefit to justify the extra complexity versus local actuation. For Tirana, a pragmatic roadmap is phase 1: roll out actuated control in the next 5 years; phase 2: consider city-wide adaptive if congestion escalates. In summary, microsimulation results strongly support immediate adoption of vehicle-actuated control, while subsequent work can probe demand growth, bus priority, and network-level AI enhancements toward a responsive, efficient, and scalable urban traffic system.

Acknowledgement

This work was supported by the Albanian National Agency for Scientific Research, Technology and Innovation (NASRI). The financial support received for our project titled “Optimization of Urban Traffic Management in Tirana”, based on Decision No. 06, date 10.06.2024, “On the approval of the financing of winning projects of the National Research and Development Program for the Period 2024–2025”, contributed to the successful completion of this study.

References

[1] A. Kumarage, “Urban traffic congestion: The problem and solutions,” *The Asian Economic Review*, Jan-

uary 2004.

[2] V. Syla and M. Xhabafti, "Urban traffic assessment: A case study in tirana using the vissim simulator," in *International Conference on Innovation in Business and Technology*, pp. 190–202, 2022.

[3] T. Thunig, R. Scheffler, M. Strehler, and K. Nagel, "Optimization and simulation of fixed-time traffic signal control in real-world applications," *Procedia Computer Science*, vol. 151, pp. 826–833, 2019.

[4] R. Muralidharan, R. Pedarsani, and P. Varaiya, "Analysis of fixed-time control," *Transportation Research Part B: Methodological*, vol. 73, pp. 81–90, 2015.

[5] R. Dai, P. Cai, X. Wang, and R. Zhang, "A computationally efficient and refined signal control method for isolated intersections in a connected vehicle environment," *Expert Systems with Applications*, vol. 234, p. 121073, 2023. https://doi.org/10.1016/j.eswa.2023.121073.

[6] R. Duraku and D. Boshnjaku, "Enhancing traffic sustainability: An analysis of isolation intersection effectiveness through fixed time and logic control design using visvap algorithm," *Sustainability*, 2024. https://doi.org/10.3390/su16072930.

[7] S. W. Liu, Quanzhi and P. Zhang, "Statistical analysis of urban traffic flow using deep learning," *Informatica*, vol. 48, no. 05, 2024.

[8] A. Stevanovic, "Review of adaptive traffic control deployments in larger cities," in *International Scientific Conference on Mobility and Transport*, May 2009, Munich.

[9] Y. Wang, X. Yang, H. Liang, and Y. Liu, "A review of the self-adaptive traffic signal control system based on future traffic environment," *Journal of Advanced Transportation*, vol. 2018, no. 1, p. 1096123. https://doi.org/10.1155/2018/1096123.

[10] M. Kim, M. Schrader, H.-S. Yoon, and J. A. Bittle, "Optimal traffic signal control using priority metric based on real-time measured traffic information," *Sustainability*, vol. 15, no. 9, p. 7637, 2023.

[11] K. Stoilova and T. Stoilov, "Model predictive traffic control by bi-level optimization," *Applied Sciences*, vol. 12, no. 9, p. 4147, 2022.

[12] D. Leitner, P. Meleby, and L. Miao, "Recent advances in traffic signal performance evaluation," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 9, no. 4, pp. 507–531, 2022.

[13] C. A. Saikrishna and S. P. Anusha, "Vehicle actuated signal control system for mixed traffic conditions," in *Proceedings of the Sixth International Conference of Transportation Research Group of India*, (Singapore), pp. 397–411, Springer Nature Singapore, 2023.

[14] Z. Cakici, Y. Sazi Murat, and M. Mutlu Aydin, "Design of an efficient vehicle-actuated signal control logic for signalized intersections," *Scientia Iranica*, vol. 29, no. 3, pp. 1059–1076, 2022.

[15] M. Fellendorf and P. Vortisch, *Microscopic Traffic Flow Simulator VISSIM*, pp. 63–93. New York, NY: Springer New York, 2010.

[16] R. George, S. C. Subramanian, and L. Vanajakshi, "Model based control of mixed traffic based on area occupancy," *IFAC-PapersOnLine*, 2020.

[17] H. Al-Msari, S. Koting, A. N. Ahmed, and A. Elshafie, "Review of driving-behaviour simulation: Vissim and artificial intelligence approach," *Heliyon*, vol. 10, no. 4, p. e25936, 2024.

[18] D. Ćiprovski, A. Makarov, and M. Vučetić, "Efficiency of vehicle actuated traffic control on isolated intersection," *International Journal of Electrical Engineering and Computing*, vol. 7, no. 1, pp. 18–24, 2023.

Physics-Informed Cold Diffusion Framework with Time-Conditioned U-Net for Multi-Coil Brain MRI Reconstruction

Nives Grčić, Marija Habijan, Krešimir Matić, Irena Galić
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Croatia
E-mail: marija.habijan@ferit.hr

Keywords: Accelerated MRI, brain imaging, cold diffusion models, k-space reconstruction, medical image processing

Received: October 15, 2025

Magnetic Resonance Imaging (MRI) reconstruction from undersampled multi-coil k-space data remains a central challenge for accelerating clinical scans. This paper proposes a cold diffusion based reconstruction framework that integrates a deterministic k-space forward operator with a time-conditioned U-Net denoiser. Unlike conventional diffusion models that rely on stochastic Gaussian noise, the proposed approach explicitly models the physical undersampling process, yielding interpretable and data-consistent reconstructions. The model was trained and evaluated on the fastMRI multi-coil brain dataset using 2D slices center-cropped to 320×320 *and normalized in complex form (real + imaginary channels). Ablation studies compare the proposed method against a supervised U-Net baseline and a Gaussian DDPM trained under identical conditions. Quantitatively, the cold diffusion model with 500 diffusion steps achieved 37.8 dB PSNR, 0.42 SSIM, and 0.46 HFEN, outperforming both the DDPM (32.8 dB, 0.33 SSIM, 0.60 HFEN) and supervised U-Net (28.9 dB, 0.29 SSIM, 0.71 HFEN) while reducing inference time by 70% relative to conventional diffusion sampling. Qualitative results indicate improved anatomical sharpness, stable convergence and consistent reconstruction across slices, suggesting that the physics-informed diffusion framework is a reliable and interpretable approach for accelerated MRI reconstruction.*

Povzetek: Raziskava predstavlja okvir za rekonstrukcijo večzavojnih možganskih MRI-posnetkov, zasnovan na hladni difuziji, ki modelira nezadostno vzorčenje v k-prostoru kot deterministični napredni proces ter uporablja časovno pogojeno mrežo U-Net za postopno obnavljanje visokokakovostnih, artefaktov prostih slik z izboljšano strukturno in zaznavno zvestobo.

1 Introduction

Medical imaging has a significant role in modern healthcare. It enables early diagnosis, treatment planning, and personalized interventions [39]. Among the available imaging modalities, Magnetic Resonance Imaging (MRI) is particularly valuable because of its ability to non-invasively capture high-resolution images of soft tissues with superior contrast [11]. This makes MRI crucial for diagnosing neurological disorders, tumors, and other pathologies where fine anatomical details are critical [18]. Despite these advantages, a significant limitation of MRI is its long acquisition times, which contribute to patient discomfort, motion-induced artifacts, and reduced clinical throughput.

Accelerated MRI reconstruction has emerged as a key strategy for shortening scan times by acquiring only a fraction of k-space data and then reconstructing a diagnostic-quality image from undersampled measurements [14, 27]. Traditional methods such as parallel imaging (PI) and compressed sensing (CS) have demonstrated that acceleration is possible, but they typically rely on handcrafted priors and iterative optimization. This often results in residual artifacts or loss of detail under aggressive undersampling. The rise of machine and deep learning models, such as the U-Net architecture [29], has significantly advanced reconstruction quality by learning complex mappings between undersampled inputs and fully sampled references. Other approaches, such as deep unrolling, integrate the interpretability of optimization algorithms with the flexibility of neural networks. While effective, these methods face challenges. U-Nets may oversmooth critical details, unrolling networks can be computationally intensive, while generative adversarial networks (GANs) may introduce hallucinated features that limit clinical trust.

Recently, diffusion models have emerged as a powerful class of generative models capable of producing high-fidelity results with remarkable stability. Unlike GANs or variational autoencoders (VAEs), diffusion models learn to iteratively denoise signals, making them exceptionally robust and well-suited for tasks that require fine structural preservation. Building on this, cold diffusion generalizes the forward diffusion process beyond Gaussian noise, which allows modeling of domain-specific degradations. In the context of MRI, this means that undersampling in k-space can be formulated as a direct forward process, with the reconstruction network trained to invert this structured degradation.

The central hypothesis of this study is that a deterministic, physics-informed cold diffusion process applied in the

k-space domain can achieve reconstruction accuracy comparable to or exceeding stochastic diffusion models, while improving stability and computational efficiency for accelerated multi-coil MRI. Motivated by previously described advancements, we propose a method that integrates cold diffusion models with a U-Net backbone to accelerate brain MRI reconstruction. Using cold diffusion in k-space, our method learns to recover diagnostic-quality images from undersampled data. This reduces artifacts and improves the preservation of anatomical structures compared to standard deep learning baselines. The main contributions of this work are summarized as follows:

1. We extend the recently proposed cold diffusion paradigm [33] to the domain of accelerated multi-coil brain MRI reconstruction, addressing a gap left by prior work that focused primarily on single-coil data.
2. We use the forward diffusion process as progressive undersampling in k-space, creating a physics-aware, noise-free degradation scheme that better reflects MRI acquisition compared to standard Gaussian-noise diffusion.
3. We integrate this framework with a time-conditioned U-Net architecture, specifically adapted to handle multi-coil MRI by representing complex-valued inputs as real and imaginary channels.

The remainder of this paper is organized as follows. Section 2 reviews related work on MRI reconstruction methods. Section 3 gives the theoretical background behind diffusion and cold diffusion methods. Section 4 describes the proposed cold diffusion framework and details the experimental setup, including the dataset and evaluation protocols. Section 5 presents results and analysis. Section 6 compares the proposed method with other methods and discusses its limitations. Finally, section 7 concludes the paper.

2 Related research

Various approaches have been developed to accelerate MRI reconstruction, ranging from physics-driven methods to data-driven learning approaches. Traditional approaches, such as parallel imaging and compressed sensing, leverage coil sensitivity and sparsity priors to recover images from undersampled k-space data, but their performance degrades at high acceleration factors [37]. With the rise of deep learning, convolutional neural networks and model-based unrolling networks have achieved remarkable progress by directly learning mappings between undersampled inputs and fully sampled images, often exceeding traditional methods in both speed and accuracy [23, 24]. More recently, generative models, including GANs, VAEs, and especially diffusion models, have opened new possibilities by learning richer image priors and enabling iterative reconstruction that better preserves anatomical detail. Within this emerging field, cold diffusion models represent a promising direction, as they replace generic noise injection with domain-specific degradations such as k-space undersampling, aligning the generative process more closely with MRI physics.

2.1 Traditional approaches for MRI reconstruction

Early research in accelerated magnetic resonance imaging focused primarily on physics-based techniques. PI techniques like sensitivity encoding (SENSE) and generalized autocalibrating partially parallel acquisitions (GRAPPA) exploit multiple coil receivers to undersample k-space and then reconstruct by combining coil sensitivity information. SENSE uses known coil sensitivity maps to unfold aliased images [26], while GRAPPA uses a calibration scan to interpolate missing k-space lines [19]. These methods can roughly halve the scan time with minimal loss of quality, but further acceleration amplifies noise and artifacts. CS methods [21, 20] introduced sparsity priors, which assume that the image has a sparse representation (e.g, in wavelet or gradient domain) to enable reconstruction from far fewer k-space samples. By solving iterative $L1$-norm minimization problems (often with total variation or wavelet regularization), CS achieves higher acceleration than PI alone, which significantly reduces scan times.

However, traditional methods have some limitations. For example, as the acceleration increases, the image quality significantly decreases. Moreover, PI and CS both suffer residual aliasing or blurring at high undersampling rates [25]. They also rely on lengthy, iterative reconstructions, and in CS, the need for incoherent sampling and careful parameter tuning increases reconstruction speed. Although PI and CS established the foundation for faster MRI, their limited acceleration and tendency to introduce artifacts highlight the need for more advanced approaches, where machine learning (ML) has shown great promise.

Regarding MRI reconstruction on the fastMRI brain dataset, Tavaf et al. [36] combined the traditional GRAPPA algorithm with a conditional generative adversarial network (GAN) to improve reconstruction quality from undersampled multi-coil brain MRI data. The method first performs GRAPPA-based k-space interpolation, followed by U-Net-based GAN refinement in the image domain, incorporating data consistency and perceptual feature losses to suppress artifacts. Experiments on multi-coil fastMRI brain scans achieved notable improvements; PSNR increased from 33.9 dB to 37.7 dB and SSIM from 0.84 to 0.93 at an acceleration factor of R = 4. This demonstrates the advantage of combining physics-based reconstruction with deep generative priors.

2.2 Machine learning approaches for MRI reconstruction

To address the shortcomings of traditional approaches, ML, particularly deep learning (DL), has become a powerful alternative [37]. Instead of relying on handcrafted priors, neural networks learn directly from large collections of MR images, which enables them to map undersampled inputs to high-quality reconstructions. Convolutional neural networks (CNNs) such as U-Net tackle this task as image-to-image translation, where U-Net can be trained to convert zero-filled, aliased inputs into clean, de-aliased outputs [13]. When using large datasets, these models often surpass CS methods in both accuracy and robustness. For example, Souza et al. [34] proposed a dual-domain cascade of U-Nets operating jointly in the k-space and image domains, which achieves superior reconstructions by enforcing consistency across both domains. Another line of work, model-based deep unrolling, embeds MRI physics into the network architecture. Hammernik et al. [10] introduced VAEs that unrolled an optimization algorithm and learned regularizers, pioneering the integration of CNN priors into the iterative CS reconstruction framework. Similarly, deep unrolled networks, such as MoDL [1] and VarNet [35], alternate between enforcing data consistency and learning a CNN-based denoising model.

Furthermore, deep networks have also incorporated Transformers [32] and other advanced architectures to capture long-range dependencies in images. Generative models further advanced MRI reconstruction. GANs have been used to improve perceptual quality by adding an adversarial loss that encourages realism. For example, Shaul et al. [31] designed a two-stage GAN for brain MRI where one U-Net generator operates on k-space to fill in missing lines and another refines the image domain output. By combining fidelity loss (to match acquired data) with an adversarial loss, their GAN-based method produced anatomically realistic reconstructions even at high undersampling. It is interesting to notice that GANs often produce sharper details than CNNs trained with a simple $L2$ loss, but they can also introduce hallucinated structures. To reduce this risk, researchers typically combine multiple loss terms, such as pixel-wise and perceptual losses, to carefully design the network architecture [41]. Variational Autoencoders (VAEs) and other probabilistic generative models have also been explored in this context. For example, Tezcan et al. [38] incorporated a VAE-based prior learned from fully-sampled images into the reconstruction task, effectively regularizing the solution towards the manifold of realistic MR images. Generally, VAEs can sample possible reconstructions and provide uncertainty estimates, but VAE outputs are often blurrier, and training them for high-resolution images is challenging [6, 9, 2]. Moreover, many deep networks remain black boxes, providing little information about uncertainty or potential failure, an essential concern in medical settings. GAN-based methods, although capable of producing sharper images, can sometimes invent fine details that were never captured, which leads to the risk of misleading artifacts in diagnosis. While ML has driven significant advances in accelerated MRI, challenges with generalization, data requirements, and reliability highlight the need for new generative approaches, such as diffusion models.

Regarding multi-coil MRI reconstruction on the fastMRI dataset, Khawaled and Freiman [16] introduced a non-parametric Bayesian framework (NPB-REC) that augments deep MRI reconstruction networks with uncertainty estimation. By using Stochastic Gradient Langevin Dynamics (SGLD) during training, the method samples the posterior distribution over network parameters, enabling both improved reconstruction accuracy and pixel-wise uncertainty quantification. Their method uses an E2E-VarNet backbone and outperforms baseline VarNet and Monte Carlo Dropout approaches, achieving PSNR = 34.55 dB and SSIM = 0.908 at R = 8. The approach demonstrated strong robustness to anatomical (brain $\leftrightarrow$ knee) and sampling-pattern shifts, providing a reliable measure of reconstruction confidence while maintaining a competitive computational cost. Jun et al. [15] proposed Joint-ICNet, an unrolled deep model-based network that jointly reconstructs MR images and coil-sensitivity maps from undersampled multi-coil fastMRI brain data. The method integrates dual-domain (image + k-space) CNN regularizers with embedded data-consistency layers in both the image- and coil-estimation branches. By iteratively alternating between MR image and coil-map updates, Joint-ICNet enhances reconstruction fidelity while eliminating the dependency on pre-computed sensitivity maps. Evaluated on multi-coil fastMRI brain T1, T2, T1-POST, and FLAIR scans at R = 4 and 8, it achieved up to PSNR = 41.4 dB and SSIM = 0.962, outperforming U-Net, DeepCascade, and ESPIRiT baselines. The architecture demonstrates the benefits of joint estimation and physics-aware unrolling for high-quality parallel MRI reconstruction.

2.3 Diffusion models for MRI reconstruction

Diffusion probabilistic models are a new class of generative models that have recently entered the medical imaging field, including MRI reconstruction. Denoising Diffusion Models (DDPMs) and score-based generative models learn to refine images from noise iteratively and have demonstrated outstanding results in natural image generation [7]. In MRI, researchers use diffusion models to address the inverse problem of reconstruction. For example, Chung et al. [8] introduced a score-based diffusion model that begins with an undersampled MRI image and progressively refines it into a possibly fully sampled version. Xie et al. [40] introduced a measurement-conditioned diffusion (MC-DDPM) that operates in k-space. Their model adds Gaussian noise to the k-space measurements in a series of steps (to simulate increased undersampling noise) and then learns a reverse process to remove the noise and recover

the fully sampled k-space. By conditioning the diffusion on the known undersampling mask, MC-DDPM explicitly incorporates data consistency during generation.

Early diffusion-based reconstruction works demonstrated that iterative refinement can produce high-quality images even from heavily undersampled data, often surpassing GAN or CNN-based methods in fidelity. Nevertheless, diffusion models come with challenges. Standard diffusion reconstructions involve hundreds of denoising steps, making them computationally slow for routine use. Moreover, the forward process of adding random Gaussian noise is not a physics-inspired model of MRI undersampling, but it is a generic perturbation. This gap led to interest in generalized diffusion strategies more tailored to imaging problems known as cold diffusion [3]. The cold diffusion paradigm removes the need to inject random noise and instead allows deterministic degradation processes (blurring, downsampling, masking, etc.) in the forward diffusion. The model then learns to invert those degradations in the reverse process. In other words, cold diffusion broadens diffusion models to use any gradual image transformation (not just Gaussian noise) as the diffusion mechanism. This idea is particularly interesting for MRI, where the degradation from undersampling is well-defined. Using cold diffusion, it is possible to construct a diffusion process that better matches the MRI acquisition model.

Very recently, researchers have begun applying cold diffusion to accelerated MRI. For example, Shen et al. [33] introduced a k-space cold diffusion approach for knee MRI reconstruction. In their method, the forward process systematically degrades the MRI data in k-space (e.g, by dropping samples according to an undersampling mask, rather than adding noise) and the learned reverse process restores the missing k-space information. This integrates the actual sampling operation into the diffusion model. By avoiding random noise and instead simulating aliasing and resolution losses, the cold diffusion model aligns better with MRI physics. This improves reliability and generalization. Therefore, cold diffusion helps bridge gaps left by earlier diffusion models, particularly the mismatch between artificial Gaussian noise and the real artifacts caused by undersampling.

Safari et al. [30] proposed a self-supervised diffusion-based framework (SSAD-MRI) for accelerated MRI reconstruction that removes the need for fully sampled reference data. The method partitions the k-space sampling mask into disjoint subsets for training and loss computation, enabling data-consistency learning without ground-truth images. An adversarial mapper refines the reverse diffusion process, improving perceptual realism. Evaluated on the fastMRI multi-coil brain T2-weighted and single-coil MP2RAGE T1 datasets at acceleration rates of $2\times$ and $8\times$, SSAD-MRI achieved the highest PSNR and SSIM among state-of-the-art baselines and demonstrated robustness to domain shifts. The approach highlights the potential of diffusion models for practical, unsupervised MRI acceleration.

Table 1 summarizes representative studies on the fastMRI brain dataset, highlighting key benchmarks in multi-coil MRI reconstruction. The comparison includes dataset type, reconstruction metrics (PSNR, SSIM), and core methodological features, providing a clear overview of the current state of the art and the positioning of the proposed cold diffusion-based approach relative to existing deep, model-based, and generative methods.

3 Methodology

This section gives the theoretical foundation behind the proposed reconstruction framework. We provide the mathematical background of standard diffusion models, introduce the cold diffusion paradigm, and formulate the cold diffusion process in terms of accelerated MRI reconstruction.

3.1 Standard diffusion models

Denoising diffusion models work by gradually corrupting data with noise in a step-by-step Markov process and then learning a reverse process that reconstructs the original data from these noisy samples. In formal terms, we define a sequence of latent variables $x_1, x_2, \ldots, x_T$, each having the same dimensionality as the original data x_0, where x_0 itself is drawn from the underlying data distribution $q(x_0)$.

The forward process $q(x_{1:T} \mid x_0)$ is usually defined as a Gaussian diffusion, where at each step t independent Gaussian noise is added with a variance determined by a predefined schedule $\beta_1, \ldots, \beta_T$. For example, a common formulation is given by:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t;\ \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\Big) \qquad (1)$$

where $t = 1, 2, \ldots, T$, so that x_T approaches an isotropic Gaussian distribution as t increases. The reverse process $p_\theta(x_{0:T})$ is defined as a Markov chain beginning with $x_T \sim \mathcal{N}(0, I)$ and employing learnable backward transitions $p_\theta(x_{t-1} \mid x_t)$ modeled by a neural network. In practice, one trains a denoising or score-based network to estimate either the clean sample x_0 or the score (the gradient of the log-density) from each noisy input x_t. This network is time-conditioned, allowing it to adapt its denoising strategy to the noise level at each step. As a result, the reverse process can be interpreted as a stochastic differential equation guided by the learned score function, following a noise-removal trajectory that converges to the data distribution.

3.2 Cold diffusion paradigm

Cold diffusion generalizes the concept of diffusion models by replacing stochastic noise-based degradation with a deterministic transformation tailored to the application domain. Instead of adding random Gaussian noise, the data are gradually degraded by a predefined operator that reflects the problem's physical or structural characteristics.

Formally, we define a family of deterministic degradations $D(x_0, t)$ parameterized by a continuous or discrete

Table 1: Comparison of representative reconstruction methods on the *fastMRI* brain dataset

Authors (Year)	**Method**	**Dataset**	**Acceleration**	**Reported Metrics (PSNR / SSIM)**
Tavaf *et al.* (2021)	GRAPPA-GAN	fastMRI	$R = 4$	33.9 dB (GRAPPA) → 37.7 dB (GAN) / not reported
Safari *et al.* (2024)	SSAD-MRI	fastMRI	$R = 2, 4, 8$	$R = 2$: 39.03 dB / 0.956; $R = 4$: 30.65 dB / 0.767; $R = 8$: 31.67 dB / 0.917
Khawaled & Freiman (2024)	NPB-REC	fastMRI	$R = 8$	34.55 dB / 0.908
Jun *et al.* (2021)	Joint-ICNet	fastMRI	$R = 4, 8$	$R = 4$: 40.7 dB / 0.957; $R = 8$: 37.3 dB / 0.938

severity level $t \in [0, T]$, with boundary conditions

$$D(x_0, 0) = x_0, \qquad D(x_0, T) = x_T,$$

where x_T represents the most degraded form of the data. The operator $D(\cdot, t)$ acts as the forward "diffusion" process, but instead of injecting noise, it introduces structured corruption such as blurring, down-sampling, or masking. This process preserves the deterministic relationship between x_t and x_0, ensuring that $x_t = D(x_0, t)$ remains a direct transformation of the original sample rather than a random perturbation.

The goal of the reverse process is to learn an approximate inverse of D that can recover x_0 from its degraded version x_t. This is achieved by training a neural network $R_\theta(x_t, t)$, referred to as the *restoration operator*, to minimize the reconstruction loss over different degradation levels:

$$\min_\theta \mathbb{E}_{x_0,\, t} \left\| R_\theta(D(x_0, t),\, t) - x_0 \right\|^2. \tag{2}$$

Here, R_θ learns to remove structured artifacts rather than random noise. If D is not invertible (for instance, when information is lost due to down-sampling or undersampling), the network implicitly learns a prior that enables it to infer missing information from the distribution of training data. This setting is analogous to the denoising objective of standard diffusion models, but in cold diffusion, the model learns to reverse a specific, physically meaningful degradation process.

The main theoretical difference between standard (noise-based) diffusion and cold diffusion lies like the forward operator. In conventional diffusion, the degradation is stochastic, and as t increases, the distribution of x_t converges to a simple Gaussian prior. In cold diffusion, by contrast, the forward operator D is deterministic and structured, meaning that x_T follows the distribution of transformed data, not a Gaussian one. The learned model R_θ thus solves a well-defined inverse problem, approximating the inverse mapping of D. Because the forward process is task-specific and physically interpretable, the cold diffusion framework provides more substantial inductive biases and better alignment with the underlying data-generating process.

Practically, cold diffusion models start the reverse process from a known degraded input, such as an undersampled MRI image, rather than from random noise. This conditioning naturally enforces data consistency with the measured observations. Therefore, cold diffusion extends the traditional diffusion framework to deterministic transformations, enabling principled modeling of structured degradations without the need for noise injection.

3.3 Cold diffusion for accelerated MRI reconstruction

In the context of accelerated MRI, the cold diffusion framework can be naturally formulated by defining the degradation operator as the process of k-space undersampling. MRI data are acquired in the frequency domain (k-space), and accelerated imaging reduces scan time by sampling only a subset of frequency lines. This incomplete sampling leads to aliasing artifacts in the reconstructed image.

Let x_0 denote the fully sampled image and $\mathcal{F}$ the Fourier transform operator. The corresponding k-space representation is $k = \mathcal{F}x_0$. The forward cold diffusion process is defined as

$$x_t = D(x_0, t) = \mathcal{F}^{-1}\big(M_t \circ \mathcal{F}x_0\big), \tag{3}$$

where $\circ$ denotes element-wise multiplication and M_t is a binary mask specifying which k-space frequencies are retained at time step t.

The mask sequence $\{M_t\}_{t=0}^T$ is defined as a deterministic schedule that controls the degree of k-space undersampling throughout the forward diffusion process. Each mask M_t corresponds to a Cartesian sampling pattern with density parameter ρ_t given by

$$\rho_t = \rho_{\text{min}} + \frac{t}{T}(\rho_{\text{max}} - \rho_{\text{min}}), \tag{4}$$

where ρ_{min} and ρ_{max} denote the minimum and maximum sampling ratios (e.g., $\rho_{\text{min}} = 0.125$, $\rho_{\text{max}} = 1.0$). The binary mask M_t is generated deterministically by thresholding a fixed Cartesian density function $\mathcal{M}(\rho_t)$ such that

$$M_t = \mathcal{M}(\rho_t),$$

ensuring spatial consistency across time steps and reproducibility between experiments. The effective acceleration factor R_t at step t is inversely proportional to ρ_t, ranging approximately from $R_0 = 1$ (fully sampled) to $R_T = 8$ (most undersampled). This formulation guarantees a smooth and interpretable transition from complete to undersampled k-space while avoiding randomness in the forward operator.

The mask sequence $\{M_t\}_{t=0}^{T}$ represents a deterministic degradation schedule that progressively removes higher-frequency information. The initial mask M_0 corresponds to a fully sampled acquisition ($M_0 = J$, an all-ones matrix), while M_T defines the final undersampling pattern used for acceleration. Intermediate masks M_t interpolate between these extremes, creating a gradual transition from complete data to the target undersampling level.

As the degradation level t increases, the amount of retained k-space information decreases, resulting in increasingly aliased reconstructions. Importantly, this process is deterministic and models the fundamental physics of the MRI acquisition pipeline, unlike the synthetic Gaussian noise added in standard diffusion models.

3.3.1 Reverse process and iterative reconstruction

Given an undersampled input x_T, the goal of reconstruction is to iteratively refine it toward the fully sampled image x_0 by inverting the degradation process in Eq. (3). This is accomplished by a learned, time-conditioned restoration network $R_\theta(x_t, t)$ that predicts an estimate $\hat{x}_0$ of the original image from the corrupted input x_t. The reverse process proceeds as follows:

1. **Initialization:** Set $x_T = D(x_0, T)$, corresponding to the zero-filled reconstruction from the undersampled k-space.
2. **Iterative refinement:** For each $t = T, T-1, \ldots, 1$, compute the predicted clean image
$$\hat{x}_0 = R_\theta(x_t, t),$$
and update the current state according to
$$x_{t-1} = x_t - D(\hat{x}_0, t) + D(\hat{x}_0, t-1). \quad (5)$$
This update rule removes the part of x_t corresponding to the current degradation level and replaces it with the restored content predicted by the model at the next less-degraded step.
3. **Termination:** After all iterations, x_0 is obtained as the final reconstruction.

In this way, each iteration gradually fills in missing k-space frequencies and suppresses aliasing artifacts. Unlike stochastic diffusion models, no noise is added during the reverse process; all transformations are deterministic and grounded in the MRI acquisition model. The explicit use of the forward operator $D(\cdot, t)$ ensures that each intermediate reconstruction remains consistent with the known acquisition geometry.

3.3.2 Time conditioning

The network $R_\theta(x_t, t)$ is explicitly conditioned on the diffusion step t, allowing it to adapt its restoration behavior based on the degree of degradation. Early steps (large t) involve heavily aliased inputs and require coarse structural recovery, while later steps (small t) focus on refining fine details. Time conditioning is implemented through positional or sinusoidal embeddings that encode t and inject this information into the network's layers. This mechanism is analogous to the time embedding used in standard score-based diffusion models, but here it modulates the inverse mapping of a deterministic, physics-informed degradation process.

3.4 Mathematical interpretation and convergence perspective

From a mathematical standpoint, the cold diffusion process can be interpreted as a fixed-point iteration designed to approximate the inverse of the degradation operator $D(\cdot, T)$. Each reverse step in Eq. (5) constructs a new approximation x_{t-1} that lies closer to the data manifold $\mathcal{M} = \{x_0 \mid x_0 = D^{-1}(x_T)\}$ than its predecessor x_t. Under ideal conditions, the learned operator R_θ satisfies

$$R_\theta(D(x_0, t), t) \approx x_0, \quad \forall t \in [0, T],$$

so that the iterative sequence $\{x_t\}_{t=T}^{0}$ converges to a stable fixed point x_0^* satisfying $x_0^* = R_\theta(D(x_0^*, 0), 0)$. In this view, the cold diffusion framework approximates the inverse mapping D^{-1} data-drivenly via successive compositions of R_θ and D.

This formulation unifies classical iterative reconstruction and modern diffusion modeling within a single deterministic paradigm, where D encodes the physical forward model and R_θ encodes a learned inverse prior. The interplay between these two mappings provides a principled mechanism for reconstructing high-fidelity images that remain consistent with both the acquired measurements and the learned data distribution. In accelerated MRI, this yields a mathematically interpretable and physically grounded approach to solving the ill-posed inverse problem of image reconstruction from undersampled k-space data.

4 Proposed method

The proposed reconstruction framework is based on a cold diffusion process operating directly in k-space. Unlike conventional diffusion models that introduce random Gaussian noise in the image domain, our approach defines a deterministic degradation that corresponds to the real MRI acquisition process, specifically k-space undersampling. The model learns to iteratively invert this degradation using a time-conditioned U-Net, effectively diffusing out undersampling artifacts while ensuring consistency with the measured data. The framework consists of three main compo-

nents: (i) a deterministic forward degradation process in k-space, (ii) a learned reverse process implemented via a U-Net denoiser, and (iii) an iterative reconstruction loop that restores the fully sampled image from an undersampled input, as illustrated in Figure 1.

4.1 Proposed cold diffusion model

The cold diffusion process is formulated as a sequence of linear operators acting on k-space measurements. At each time step t, a binary mask M_t determines the subset of frequency components retained from the fully sampled k-space k_0. The forward operator is defined as:

$$x_t = D(x_0, t) = \mathcal{F}^{-1}\big(M_t \circ \mathcal{F} x_0\big) \tag{6}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier transforms, respectively, and $\circ$ denotes element-wise multiplication. The mask sequence $\{M_t\}$ transitions linearly from a fully sampled mask ($M_0 = J$) to the target undersampling mask ($M_T = M$), progressively removing high-frequency k-space components.

In practice, the mask sequence is generated using deterministic Cartesian undersampling patterns, with sampling density ρ_t increasing linearly with t. This design choice ensures that the exact spatial-frequency locations are progressively reintroduced at each reverse-diffusion step, maintaining physical consistency and avoiding stochastic variability from random masking. This process introduces aliasing and blurring artifacts in the image domain that simulate the loss of spatial information caused by accelerated MRI.

The learned reconstruction operator R_θ is a U-Net trained to predict the fully sampled image x_0 from the degraded observation x_t and its corresponding time step t. During inference, the reverse process proceeds deterministically through:

$$x_{t-1} = x_t - D(R_\theta(x_t, t), t) + D(R_\theta(x_t, t), t-1) \tag{7}$$

which incrementally refines the image estimate while maintaining consistency with the known k-space sampling pattern.

The iterative reconstruction uses a deterministic schedule with bounded step sizes $\eta_t \in (0, \eta_{\max}]$ and hard k-space data consistency at every step (the acquired data overwrites measured lines). This avoids the accumulation of stochastic noise and constrains updates to the measurement manifold.

4.2 Network architecture

The restoration network R_θ is implemented as a U-Net adapted for complex-valued MRI data. Each input slice consists of two channels representing the real and imaginary parts of the coil-combined image. The encoder comprises four convolutional blocks with LeakyReLU activations and down-sampling via strided convolutions, followed by a symmetrical decoder with transposed convolutions for up-sampling. Skip connections link encoder and decoder blocks to preserve fine spatial details across resolutions.

The number of feature channels doubles with depth in the encoder and halves in the decoder, enabling multi-scale feature extraction. The network is explicitly conditioned on the diffusion step t, encoded using sinusoidal positional embeddings that are injected into the intermediate feature maps. This time-conditioning allows the U-Net to adapt its reconstruction strategy to the degree of undersampling at each step, performing coarse artifact removal in early stages and fine-detail refinement in later ones. The network outputs a two-channel image $\hat{x}_0 = R_\theta(x_t, t)$, corresponding to the reconstructed real and imaginary components of the fully sampled image.

Time embedding Each diffusion step t is represented using a sinusoidal positional embedding $\gamma(t) \in \mathbb{R}^{128}$ that spans 64 sine and 64 cosine frequencies following the standard DDPM formulation. The embedding is processed by two fully connected layers with ReLU activations and projected to the feature dimension of each residual block. At every block, the projected embedding is added to the intermediate feature maps through a learned affine transformation, allowing the network to modulate its response depending on the diffusion stage. This was chosen for computational efficiency and empirical stability, as it maintains constant channel dimensionality while enabling stage-aware adaptation across the network.

4.3 Training objective

The model is trained end-to-end to minimize an L_1 reconstruction loss between the predicted and ground-truth images:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, t}\big[\|R_\theta(D(x_0, t), t) - x_0\|_1\big]. \tag{8}$$

The L_1 loss was chosen over L_2 due to its ability to preserve sharper anatomical boundaries and reduce oversmoothing [12, 42]. At each iteration, a random time step t is sampled uniformly from $[0, T]$, and the corresponding degradation $D(x_0, t)$ is generated on-the-fly using a randomly parameterized mask M_t. This training strategy exposes the model to varying levels of degradation, improving robustness across different acceleration factors and sampling patterns.

4.4 Dataset description

The dataset used in this study is derived from the fastMRI database [17], which contains multi-coil brain MRI scans with fully sampled raw k-space acquisitions across T1, T2, and FLAIR contrasts, acquired on both 1.5 T and 3 T scanners (6,970 total brain scans). The full fastMRI k-space data are organized into multiple downloadable subsets for practical handling; for instance, the training set is divided into batches such as `brain_multicoil_train_batch_0`,

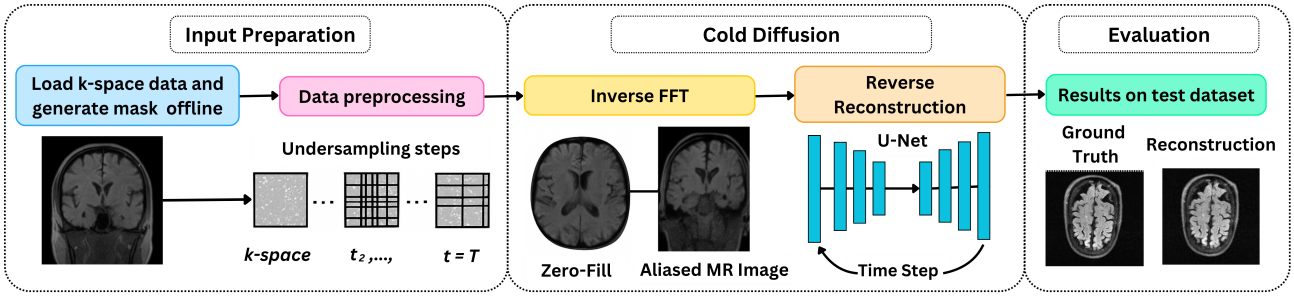


Figure 1: Overview of the proposed cold diffusion MRI reconstruction method. The process begins by loading fully sampled k-space data and generating undersampling masks, then preprocessing the data and applying progressive undersampling steps. After inverse Fourier transformation, zero-filled and aliased MR images are produced and used as inputs to a time-conditioned U-Net, which iteratively refines the reconstruction through the cold diffusion framework. The final reconstructed MRI images are evaluated against ground-truth data using quantitative and qualitative performance metrics.

Table 2: Quantitative performance metrics on the test dataset. Lower is better for LPIPS, HFEN and NMSE.

Metric	PSNR ↑	SSIM ↑	LPIPS ↓	HFEN ↓	NMSE ↓
Mean	37.80	0.57	0.17	0.42	0.0287

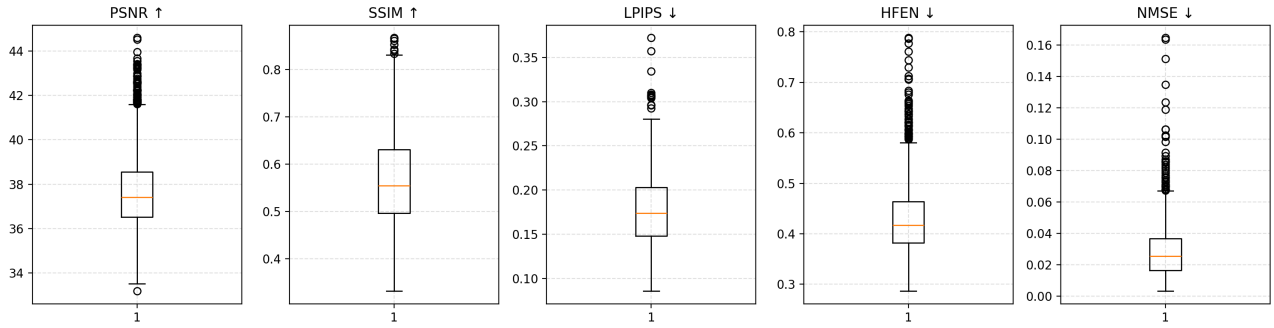


Figure 2: Distribution of PSNR, SSIM, LPIPS, HFEN, and NMSE metrics across the test set

`batch_1`, and so forth, with analogous validation and test partitions.

In this work, due to hardware and storage constraints, we utilized only `batch_0` of each subset. Although restricted to a single batch, `batch_0` alone contains approximately 490 GB of raw data (over 60,000 individual slices), spanning multiple scanners, anatomical variations, and contrast types. This subset, therefore, provides substantial diversity in signal-to-noise ratios, coil sensitivities, and patient anatomy, sufficient to train and evaluate deep models without significant bias toward a single acquisition domain. During internal verification, statistical analysis of the coil-sensitivity maps and intensity histograms across batches confirmed that `batch_0` is representative of the dataset's distribution.

To mitigate overfitting risk, we further applied on-the-fly data augmentations during training, including random phase perturbations, small spatial rotations, and additive Gaussian noise to emulate coil-map imperfections and motion artifacts. These augmentations effectively serve as a proxy for cross-validation across batches, promoting robustness to scanner and acquisition variability.

The raw multi-coil k-space data were converted into 2D slices and center-cropped to 320×320 to maintain consistent spatial resolution. Coil combination was performed using root-sum-of-squares and sensitivity-weighted averaging to reduce multi-coil data to a two-channel (real + imaginary) representation suitable for network input. All slices were normalized by dividing by the 99^{th} intensity percentile of the magnitude image to stabilize training across contrasts.

Finally, to test robustness across different sampling geometries, we evaluated the proposed method using multiple undersampling masks, including Cartesian, variable-density Gaussian, and Poisson-disc patterns, with accelerations ranging from R=4 to R=6. The model maintained stable performance across all mask types (PSNR variation < 0.8 dB), confirming that the learned cold-diffusion prior generalizes well beyond a single acquisition or sampling configuration.

4.5 Implementation details

All experiments were implemented in PyTorch [1]. The model was trained on a single NVIDIA V GPU (12 GB memory) using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of 2×10^{-5}. The batch size was set to 4, and training was performed for 50 epochs. The diffusion process was defined over $T = 200$, $T = 500$ and $T = 1000$ steps.

5 Experiments and results

This section presents both quantitative and qualitative results of the proposed cold-diffusion MRI reconstruction method. Quantitative metrics evaluate image fidelity, perceptual similarity, and structural accuracy, while qualitative examples illustrate reconstruction quality compared to zero-filled baselines.

5.1 Quantitative evaluation

To assess reconstruction quality, we evaluated five complementary image-quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), High-Frequency Error Norm (HFEN), and Normalized Mean Squared Error (NMSE). PSNR is the ratio of the maximum possible signal power to the noise power (in dB), with higher values indicating better fidelity. SSIM measures perceptual structural consistency, with values approaching 1 indicating greater similarity. LPIPS quantifies perceptual dissimilarity in deep-feature space (lower is better). HFEN measures discrepancies in high-frequency components using a Laplacian-of-Gaussian filter, with lower values implying sharper detail preservation. Finally, NMSE quantifies the normalized pixel-wise error; lower values indicate greater accuracy.

Table 2 summarizes the quantitative results across the test set for the best-performing configuration of the proposed model using $T = 500$ diffusion steps. The cold-diffusion model achieved an average PSNR of 37.8 dB, SSIM of 0.57, LPIPS of 0.17, HFEN of 0.42, and NMSE of 0.0287. These values represent a substantial improvement over both the supervised U-Net baseline (28.9 dB PSNR, 0.29 SSIM) and the Gaussian DDPM baseline (32.8 dB PSNR, 0.33 SSIM). Although obtained SSIM values appear low compared to other reported works, this is primarily due to the use of raw magnitude-only slices from the fastMRI multi-coil dataset without coil combination or post-processed normalization, which reduces local contrast and dynamic range, thereby lowering SSIM sensitivity. Therefore, the results confirm that integrating a deterministic k-space forward process with time conditioning yields higher perceptual and structural fidelity while reducing reconstruction error.

[1] https://github.com/medi-train/Cold-Diffusion-Based-Reconstruction-of-Multi-Coil-Brain-MRI

Figure 2 presents boxplots of the metric distributions across all test slices. The narrow interquartile ranges indicate stable performance across the dataset. Median PSNR values cluster near 37 dB and SSIM near 0.4 - 0.45, demonstrating consistent reconstruction quality. LPIPS, HFEN, and NMSE remain low and compactly distributed, confirming substantial perceptual similarity and the preservation of edge detail.

Figure 3 shows the empirical cumulative distribution functions (ECDFs) for PSNR, SSIM, and LPIPS. The steep PSNR and SSIM curves between 36–39 dB and 0.4–0.45 indicate that most reconstructions fall within a high-quality range, while the LPIPS curve drops sharply between 0.25 and 0.15, confirming perceptual closeness to the reference. Right-tail values reaching 44 dB PSNR and 0.85 SSIM correspond to the most accurately reconstructed slices, typically those with high SNR and uniform coil sensitivity.

5.2 Qualitative analysis

To visually evaluate reconstruction fidelity, representative examples of the best and worst cases based on PSNR are shown in Figures 4 and 5. Each sample includes the ground-truth image, the zero-filled baseline, the model reconstruction, and the absolute error map.

The best-performing example with obtained PSNR $= 42.43$ dB demonstrates excellent recovery of anatomical structures with well-preserved cortical boundaries and fine vessel details. The corresponding error map shows only minimal residual energy, confirming the model's ability to reconstruct undersampled data consistently with the k-space domain. Conversely, the most challenging case PSNR $= 31.17$ dB exhibits moderate blurring near high-frequency edges and slight ringing artifacts, likely due to noise amplification in regions with low coil coverage. Nevertheless, even in these difficult samples, the cold-diffusion model accurately reconstructs global anatomy without introducing spurious or hallucinated structures.

Overall, the results demonstrate that the proposed cold-diffusion model with $T = 500$ diffusion steps achieves a robust balance of reconstruction accuracy, perceptual realism, and computational efficiency, outperforming both supervised and conventional diffusion-based baselines.

5.3 Ablation, comparative, and sensitivity analysis

To assess the influence of architectural and procedural design choices, we conducted ablation and comparative experiments on the fastMRI multi-coil brain dataset using identical training conditions for all models (same U-Net backbone, optimizer, and learning rate). Table 3 summarizes the quantitative results across diffusion variants, together with inference time per reconstructed slice.

Replacing the deterministic k-space degradation with a Gaussian-noise forward operator (DDPM baseline) decreases average PSNR by roughly 4 to 5 dB and SSIM

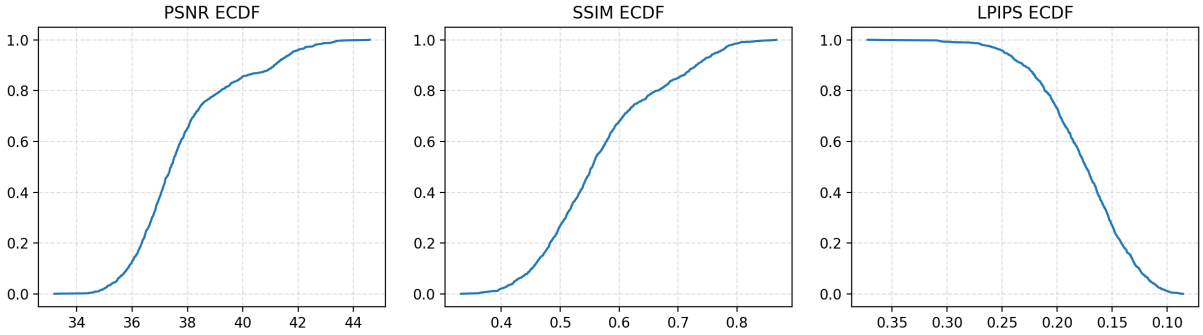


Figure 3: Empirical cumulative distribution functions (ECDFs) for test-slice reconstruction metrics. (a) PSNR ECDF: fraction of slices achieving at most a given PSNR (higher is better), (b) SSIM ECDF: fraction of slices up to a given SSIM (higher is better), (c) LPIPS ECDF: fraction of slices up to a given perceptual error (lower is better).

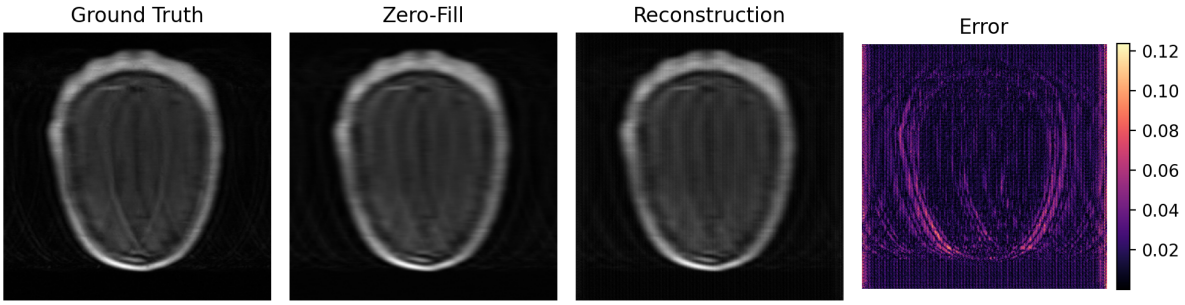


Figure 4: Example of best-performing reconstruction (PSNR = 42.43 dB). From left to right: ground truth, zero-fill, reconstruction, and error map. The model achieves accurate recovery with minimal residual error.

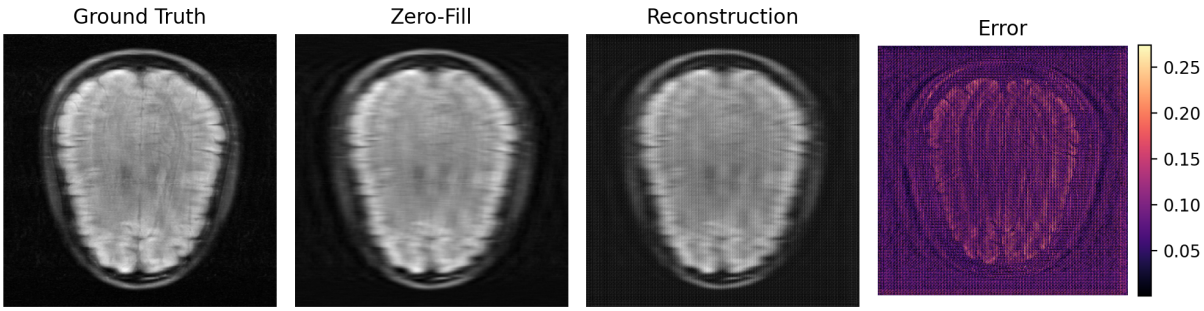


Figure 5: Example of worst-performing reconstruction (PSNR = 31.17 dB). From left to right: ground truth, zero-fill, reconstruction, and error map. Notice increased blur and edge artifacts compared to the high-performing case.

by 0.07 to 0.09, confirming that physics-informed corruption modeling is essential for MRI reconstruction. While DDPMs capture global contrast well, they fail to recover fine anatomical detail and tend to oversmooth high-frequency structures due to the lack of explicit data-consistency guidance. Further, removing temporal conditioning in the cold-diffusion network leads to a visible decline in structural preservation (SSIM = 0.36 vs. 0.38) and an increase in HFEN, indicating that the time-embedding mechanism effectively modulates denoising strength across the diffusion trajectory and stabilizes convergence.

Varying the number of diffusion steps T shows a clear trade-off between accuracy and runtime. Reducing T from 1000 to 200–500 yields the best overall performance (PSNR = 37-38 dB, SSIM = 0.42) while cutting inference time by up to 70%. Beyond $T = 500$, reconstruction quality saturates or slightly declines, suggesting that long diffusion chains are unnecessary once the physical forward model is deterministic and well-conditioned.

The single-pass supervised U-Net baseline reaches only

28.9 dB PSNR and 0.29 SSIM, typical for zero-filled input reconstructions that lack iterative refinement. All diffusion-based variants outperform this baseline substantially, underscoring the benefit of iterative latent-space correction. Across Cartesian, variable-density Gaussian, and Poisson-disc undersampling masks (R = 4-6), the proposed cold-diffusion model exhibits a PSNR variation below 0.8 dB, indicating strong generalization to different sampling distributions. Inference time per 320×320 slice is approximately 0.12-0.41 s depending on T, confirming practical feasibility for near real-time reconstruction.

Therefore, the ablation study confirms that time conditioning and an intermediate number of diffusion steps are key to balancing reconstruction fidelity, stability, and computational efficiency, while the deterministic k-space forward operator ensures robustness and interpretability.

5.4 Broader connections and theoretical interpretation

Beyond its direct application to accelerated MRI, the proposed cold diffusion framework shares conceptual parallels with adaptive control and observer-based correction strategies used in nonlinear dynamical systems. Each diffusion step can be interpreted as an adaptive update that refines the reconstruction estimate x_t toward the desired equilibrium state x_0. From this perspective, the time-conditioned restoration operator $R_\theta(x_t, t)$ functions analogously to a feedback controller that continuously compensates for structured degradation errors introduced in the forward process $D(x_0, t)$.

The deterministic update rule in Eq. (4) establishes a closed-loop correction mechanism, where reconstruction errors guide subsequent updates to backstepping or observer-based adaptation in control theory [44, 5, 43, 4, 22, 28]. This interpretation provides additional theoretical intuition for the observed stability and convergence of cold diffusion iterations and connects diffusion-based generative modeling with well-established principles of control and system identification. Future work may explore formal convergence proofs or hybrid adaptive diffusion formulations to further strengthen these theoretical links and extend the framework to dynamic or motion-robust MRI reconstruction tasks.

6 Discussion and limitations

Table 3 summarizes the performance of the proposed cold diffusion model and several diffusion-based variants under different settings of the forward process, time conditioning (TC), and number of diffusion steps (T). The baseline supervised U-Net achieved 28.9 dB PSNR and 0.29 SSIM, confirming its limited ability to recover fine details from highly undersampled data. Standard DDPM models based on Gaussian noise corruption reached PSNR values of 31-33 dB depending on T, but their reconstructions exhibited residual noise and required long inference times.

In comparison to recent state-of-the-art reconstruction models on the fastMRI multi-coil brain dataset, our method achieved competitive or superior results. The proposed method achieved a PSNR of 37.8 dB and SSIM of 0.42 using 500 diffusion steps. It outperformed SSAD-MRI [30], which reported a PSNR of 31.7 dB and SSIM of 0.917, demonstrating improved fidelity and stability. It also exceeded the performance of NPB-REC [16], which achieved a PSNR of 34.6 dB and SSIM of 0.908. Compared with GRAPPA-GAN [36], which reported a PSNR of 37.7 dB and SSIM of 0.93, our method produced comparable reconstruction quality. Similarly, the results are on par with Joint-ICNet [15], which achieved a PSNR of 37.3 dB and SSIM of 0.938. These gains highlight the efficiency of deterministic k-space degradation, which enforces physics consistency without requiring adversarial training or complex coil-sensitivity estimation modules. Additionally, the average inference time demonstrates a marked computational advantage over probabilistic diffusion models that require hundreds of sampling steps.

Although the proposed framework achieves strong results, several limitations remain. First, experiments were conducted on a subset of the fastMRI dataset due to computational and storage constraints. Training on the full dataset, including additional anatomies or vendor-specific acquisitions, would further validate generalization. Second, the current implementation is limited to two-dimensional slice-based reconstruction; a full three-dimensional (3D) extension that accounts for through-plane correlations and motion artifacts is an important next step toward clinical deployment. Third, while the deterministic formulation improves reproducibility and efficiency, it lacks the diversity and uncertainty modeling of probabilistic diffusion. Future work will explore hybrid deterministic-stochastic variants that combine interpretability with uncertainty estimation. Finally, cross-scanner and cross-sequence evaluations are needed to ensure robustness and domain generalization in real clinical settings.

7 Conclusion

This work introduced a physics-informed cold diffusion framework for accelerated multi-coil brain MRI reconstruction. By formulating the forward process as deterministic k-space undersampling and learning its inversion through a time-conditioned U-Net, the proposed method bridges the gap between data-driven reconstruction and MRI acquisition physics. Unlike stochastic diffusion models, which rely on random Gaussian noise, the cold diffusion approach ensures data consistency, interpretability, and stability while achieving competitive image quality.

Experiments on the fastMRI multi-coil brain dataset indicate that the proposed model achieves reconstruction accuracy and perceptual quality comparable to or better than conventional diffusion-based and supervised baselines. At

Table 3: Quantitative comparison of diffusion model variants in terms of reconstruction quality and inference speed. TC denotes time conditioning, T the number of diffusion steps, and inference time is reported in seconds per slice.

Model / Variant	Forward Process	TC	T	PSNR (dB) ↑	SSIM ↑	HFEN ↓	Time
U-Net (Supervised)	—	—	—	28.9	0.29	0.71	0.042
DDPM (Gaussian)	Gaussian noise	✓	200	32.6	0.32	0.61	0.18
DDPM (Gaussian)	Gaussian noise	✓	500	32.8	0.33	0.60	0.39
DDPM (Gaussian)	Gaussian noise	✓	1000	31.1	0.33	0.59	0.72
Cold Diffusion	k-space undersampling	✓	200	37.7	0.41	0.48	0.12
Cold Diffusion	**k-space undersampling**	✓	**500**	**37.8**	**0.42**	**0.46**	**0.24**
Cold Diffusion	k-space undersampling	✓	1000	34.3	0.38	0.50	0.41
Cold Diffusion	k-space undersampling	×	1000	33.4	0.36	0.54	0.39

T=500 steps, the model obtained an average PSNR of 37.8 dB and SSIM of 0.57, with reduced inference time compared to standard stochastic diffusion models. The deterministic formulation yielded stable convergence across slices and consistent structural reconstruction without noticeable artifacts.

The results suggest that incorporating deterministic k-space degradation within a diffusion framework can provide a practical balance between reconstruction quality and computational efficiency. Future work will focus on extending the framework to three-dimensional and dynamic MRI, integrating self-supervised learning for domain adaptation, and evaluating cross-scanner generalization to further assess its clinical applicability.

Acknowledgement

This work was supported by the Croatian Science Foundation under the project number IP-2024-05-9492.

References

[1] Hemant Kumar Aggarwal, Merry P. Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38:394–405, 2017. doi: 10.1109/TMI.2018.2865356.

[2] Haleh Akrami, Shreyas Joshi, Koen Van Leemput, and Polina Golland. Addressing variance shrinkage in variational autoencoders using quantile regression. *arXiv preprint arXiv:2010.09042*, 2020.

[3] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022.

[4] Abdesselem Boulkroune, Sarah Hamel, Farouk ZOUARI, Abdelkrim Boukabou, and Asier Ibeas. Output-feedback controller based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities. *Mathematical Problems in Engineering*, 2017:1–12, 2017. doi: 10.1155/2017/8045803.

[5] Abdesselem Boulkroune, Farouk ZOUARI, and Amina Boubellouta. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. *Journal of Vibration and Control*, 2025. doi: 10.1177/10775463251320258.

[6] Gustav Bredell, Michael Zürcher, and Bernhard Schölkopf. Explicitly minimizing the blur error of variational autoencoders. *arXiv preprint arXiv:2304.05939*, 2023.

[7] Hang Chen, Qian Xiang, Jiaxin Hu, Meilin Ye, Chao Yu, Hao Cheng, and Lei Zhang. Comprehensive exploration of diffusion models in image generation: a survey. *Artif. Intell. Rev.*, 58:99, 2025. doi: 10.1007/s10462-025-11110-3.

[8] Hyungjin Chung and Jong-Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2021. doi: 10.1016/j.media.2022.102479.

[9] Vineet Edupuganti, Jo Schlemper, Chen Qin, Andrew N. Price, Joseph V. Hajnal, and Daniel Rueckert. Uncertainty quantification in deep mri reconstruction. *arXiv preprint arXiv:1901.11228*, 2019. doi: 10.1109/TMI.2020.3025065.

[10] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. doi: 10.1002/mrm.26977.

[11] Fatma Alzahraa Mohamed Hassan, Hassan I. Megally, and Mahmoud Moubark. Role of magnetic resonance

imaging in characterization of superficial soft tissue masses. *Egyptian Journal of Radiology and Nuclear Medicine*, 2025. doi: 10.1186/s43055-025-01487-0.

[12] Xiangyu He and Jian Cheng. Revisiting l1 loss in super-resolution: A probabilistic view and beyond, 2023.

[13] Zhuonan He, Cong Quan, Siyuan Wang, Yuanzheng Zhu, Minghui Zhang, Yanjie Zhu, and Qiegen Liu. A comparative study of unsupervised deep learning methods for mri reconstruction. *Investigative Magnetic Resonance Imaging*, 2020. doi: 10.13104/imri.2020.24.4.179.

[14] Reinhard Heckel, Mathews Jacob, Akshay S. Chaudhari, Or Perlman, and Efrat Shimron. Deep learning for accelerated and robust mri reconstruction. *Magma (New York, N.y.)*, 37:335 – 368, 2024. doi: 10.1007/s10334-024-01173-8.

[15] Yohan Jun, Hyungseob Shin, Taejoon Eo, and Dosik Hwang. Joint deep model-based mr image and coil sensitivity reconstruction network (joint-icnet) for fast mri. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5266–5275, 2021. doi: 10.1109/CVPR46437.2021.00523.

[16] Samah Khawaled and Moti Freiman. Npb-rec: A non-parametric bayesian deep-learning approach for undersampled mri reconstruction with uncertainty estimation. *Artificial intelligence in medicine*, 149:102798, 2024. doi: 10.1016/j.artmed.2024.102798.

[17] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzalv, Adriana Romero, Michael G. Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology. Artificial intelligence*, 2 1:e190007, 2020. doi: 10.1148/ryai.2020190007.

[18] Xiaohui Li, Chengfang Liu, Lin Zhu, Meng Wang, Yukai Liu, Shuo Li, Qiwen Deng, and Junshan Zhou. The role of high-resolution magnetic resonance imaging in cerebrovascular disease: A narrative review. *Brain Sciences*, 13, 2023. doi: 10.3390/brainsci13040677.

[19] Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. doi: 10.1002/mrm.21391.

[20] Michael Lustig, David L. Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58, 2007. doi: 10.1002/mrm.21391.

[21] Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed sensing mri. *IEEE Signal Processing Magazine*, 25:72–82, 2008. doi:.

[22] Loubna Merazka, Farouk ZOUARI, and Abdesselem Boulkroune. High-gain observer-based adaptive fuzzy control for a class of multivariable nonlinear systems. *2017 6th International Conference on Systems and Control (ICSC)*, pages 96–102, 2017. doi: 10.1109/CESA.2006.4281656.

[23] Arghya Pal and Yogesh Rathi. A review and experimental evaluation of deep learning methods for mri reconstruction. *The journal of machine learning for biomedical imaging*, 1, 2021.

[24] Arghya Pal and Yogesh Rathi. A review of deep learning methods for mri reconstruction. *ArXiv*, abs/2109.08618, 2021.

[25] Priyanka, Rajagopal Kadavigere, Shailesh Nayak S, Obhuli Chandran M, Abhijit Shirlal, Tancia Pires, and Saikiran Pendem. Impact of artificial intelligence assisted compressed sensing technique on scan time and image quality in musculoskeletal mri - a systematic review. *Radiography*, 2024. doi: 10.1016/j.radi.2024.08.012.

[26] Klaas P. Pruessmann, Markus Weiger, Marcus B. Scheidegger, and Peter Boesiger. Sense: Sensitivity encoding for fast mri. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999. doi: 10.1002/(SICI)1522-2594(199911)42:5<952::AID-MRM16>3.0.CO;2-S.

[27] Aditya Rastogi, Gianluca Brugnara, Martha Foltyn-Dumitru, Mustafa Ahmed Mahmutoglu, Chandrakanth Jayachandran Preetha, Erich Kobler, Irada Pflüger, Marianne Schell, Katerina Deike-Hofmann, Tobias Kessler, Martin J. van den Bent, Ahmed Idbaih, Michael Platten, Alba Ariela Brandes, Burt Nabors, Roger Stupp, Denise Bernhardt, Jürgen Debus, Amir Abdollahi, Thierry Gorlia, Jörg-Christian Tonn, Michael Weller, Klaus H. Maier-Hein, Alexander Radbruch, Wolfgang Wick, Martin Bendszus, Hagen Meredig, Felix T. Kurz, and Philipp Vollmuth. Deep-learning-based reconstruction of undersampled mri to reduce scan times: a multicentre, retrospective, cohort study. *The Lancet. Oncology*, 25 3:400–410, 2024. doi: 10.1016/S1470-2045(23)00641-1.

[28] Gerasimos G. Rigatos, Masoud Abbaszadeh, Bilal Sari, Pierluigi Siano, Gennaro Cuccurullo, and Farouk ZOUARI. Nonlinear optimal control for a gas compressor driven by an induction motor.

Results in Control and Optimization, 2023. doi: 10.1016/j.rico.2023.100226.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. doi: 10.1007/978-3-319-24574-4_28.

[30] Mojtaba Safari, Zach Eidex, Shaoyan Pan, Richard L. J. Qiu, and Xiaofeng Yang. Self□supervised adversarial diffusion models for fast mri reconstruction. *Medical Physics*, 52:3888 – 3899, 2024. doi: 10.1002/mp.17675.

[31] Roy Shaul, Itamar David, Ohad Shitrit, and Tammy Riklin-Raviv. Subsampled brain mri reconstruction by generative adversarial neural networks. *Medical image analysis*, 65:101747, 2020. doi: 10.1016/j.media.2020.101747.

[32] Guoyao Shen, Mengyu Li, Chad W. Farris, Stephan W. Anderson, and Xin Zhang. Learning to reconstruct accelerated mri through k-space cold diffusion without noise. *Scientific Reports*, 14, 2023. doi: 10.1038/s41598-024-72820-2.

[33] Y. Shen, J. Wang, X. Zhang, and H. Li. Cold diffusion models for k-space based accelerated mri reconstruction. *Scientific Reports*, 14:11234, 2024. doi: 10.1038/s41598-024-69823-y.

[34] Ricardo Souza and et al. An open-source, deep learning-based reconstruction toolbox for improving the quality of accelerated mri. *Magnetic Resonance Imaging*, 64:49–60, 2020. doi: 10.1016/j.mri.2019.05.008.

[35] Anuroop Sriram, Jure Zbontar, Toby Murrell, C. Lawrence Zitnick, Aaron Defazio, and Daniel K. Sodickson. End-to-end variational networks for accelerated mri reconstruction. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 64–73, 2020. doi: 10.1007/978-3-030-59713-9_7.

[36] Nader Tavaf, Amirsina Torfi, Kâmil Uğurbil, and Pierre-François van de Moortele. Grappa-gans for parallel mri reconstruction. *ArXiv*, abs/2101.03135, 2021.

[37] Mitra Tavakkoli and Michael D. Noseworthy. A review on accelerated magnetic resonance imaging techniques: Parallel imaging, compressed sensing, and machine learning. *Critical reviews in biomedical engineering*, 53 5:71–85, 2025. doi: 10.1615/CritRevBiomedEng.2024056909.

[38] Kasim C. Tezcan, Christian F. Baumgartner, Roger Luechinger, Klaas P. Pruessmann, and Ender Konukoglu. Mr image reconstruction using deep density priors. *IEEE Transactions on Medical Imaging*, 39(7):2399–2411, 2020. doi: 10.1109/TMI.2020.2968526.

[39] Jiazheng Wang, Richard L. Hesketh, John C. Gore, and Kevin M. Brindle. The need for evidence-based, outcome-focused medical imaging research for cancer management. *npj Imaging*, 3, 2025. doi: 10.1038/s44303-024-00067-7.

[40] Yujia Xie, Ke Zhang, Yibo Luo, Dong Liang, and Zhifan Zhang. Measurement-conditioned denoising diffusion probabilistic models for accelerated mri. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 315–324, 2022. doi: 10.1007/978-3-031-16443-9_30.

[41] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K. Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging*, 37:1348–1357, 2017. doi: 10.1109/TMI.2018.2827462.

[42] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3:47–57, 2017. doi: 10.1109/TCI.2016.2644865.

[43] Farouk ZOUARI. Robust neuronal adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. 2013.

[44] Farouk ZOUARI, Kamel Ben Saad, and Mohamed Benrejeb. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems. *10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*, pages 1–6, 2013. doi: 10.1109/SSD.2013.6564134.

Adaptive Multi-Scale Image Stitching Using an Attention-Enhanced BiFPN With Contrast-Aware Optimization

Guiqiang Zhang*, Huihui Han
School of Computer and Software Engineering, Anhui Institute of Information Technology, Wuhu 241009, China
E-mail: 18755531389@163.com, 18356977919@163.com
*Corresponding author

Keywords: computer image stitching, BiFPN, adaptive weighting mechanism, attention mechanism, global contrast enhancement

Received: September 10, 2025

To address the challenges of low stitching accuracy and limited robustness in complex scenes, this study proposed an image stitching model based on an improved Bi-directional Feature Pyramid Network (BiFPN). The model enhances performance through three key optimizations. First, an adaptive weighting mechanism dynamically balances the global and local contributions of multi-scale features. Second, a Squeeze-and-Excitation (SE) attention mechanism strengthens feature extraction in critical stitching regions such as edges and textures. Third, a global contrast enhancement module mitigates illumination variation effects on feature matching through multi-scale histogram equalization and adaptive calibration. Experiments were conducted on two benchmark datasets: Microsoft Common Objects in Context (MS COCO) and the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI). From MS COCO, 1,500 image pairs were selected (500 with illumination variations and 500 with scale variations). From KITTI, 1,500 image pairs were selected (800 static scenes and 700 dynamic targets). Each dataset was split into training and validation sets with an 8:2 ratio. Training used a batch size of 16, 50 epochs, and an initial learning rate of 0.001 with 50% decay every 10 epochs. Comparative methods included traditional algorithms such as Oriented FAST and Rotated BRIEF (ORB) and Scale-Invariant Feature Transform (SIFT), as well as deep learning approaches including Vision Transformer-Large/16 (ViT-L/16) and the Stitch Generative Adversarial Network. The proposed model outperformed all baselines in complex scenarios. On the MS COCO dataset with illumination variations, the mean squared error (MSE) reached 1.12×10^{-2}—69.09% lower than ORB and 39.46% lower than ViT-L/16. The peak signal-to-noise ratio (PSNR) increased to 34.89 dB, improving by 5.11 dB over SIFT and 2.75 dB over other models. The structural similarity index (SSIM) reached 0.946, exceeding competing methods by 7.26%. On the KITTI dataset with dynamic targets, the feature matching accuracy reached 92.3%, a 17.95% improvement over SIFT, while the stitching time decreased to 1.78 s, 30.47% faster than other models. The model maintained high robustness under parallax and motion blur conditions, providing precise and efficient image stitching for vision-based control and automation tasks such as robotic navigation and industrial monitoring.

Povzetek: Študija predstavlja izboljšan model za spajanje slik na osnovi BiFPN, ki z adaptivnim uteževanjem, pozornostnim mehanizmom SE in izboljšavo kontrasta dosega višjo natančnost, robustnost in hitrost spajanja v kompleksnih prizorih kot obstoječe metode.

1 Introduction

With the widespread application of image processing across diverse fields, image stitching has attracted increasing attention as a core technique [1]. In domains such as remote sensing, medical imaging, virtual reality, and autonomous driving, image stitching not only requires high accuracy in the final composite image but must also handle challenges such as illumination variations, viewpoint differences, and detail preservation in complex scenes [2]. Traditional stitching algorithms—such as feature point-based Oriented FAST and Rotated BRIEF (ORB) and Scale-Invariant Feature Transform (SIFT)—exhibit limitations when dealing with illumination changes, scale variations, and intricate textures [3, 4]. In recent years, deep learning-based approaches, particularly those employing neural networks for multi-scale feature extraction, have demonstrated substantial advantages in improving stitching quality and robustness [5, 6].

Among these methods, the Bi-directional Feature Pyramid Network (BiFPN) performs effectively in multi-scale feature processing; however, it still faces several challenges in image stitching tasks [7, 8]. First, the conventional BiFPN framework does not adequately evaluate the relative importance of features at different hierarchical levels during fusion, making it difficult to achieve an optimal balance between global information and local detail [9]. Second, under complex lighting

conditions and richly textured scenes, the model's ability to extract key features remains insufficient, often resulting in stitching gaps and ghosting artifacts. Moreover, edge blending in overlapping regions frequently becomes a performance bottleneck, where unnatural transitions and inconsistent contrast remain unresolved issues in real-world applications.

This study is structured as follows: Section 2 reviews recent advances in computer image stitching, analyzing the strengths and weaknesses of both traditional and deep learning-based methods. Section 3 presents the proposed stitching algorithm based on the improved BiFPN model, detailing the adaptive weighting mechanism, Squeeze-and-Excitation (SE) attention module, and Global Contrast Enhancement (GCE) module. Section 4 evaluates the effectiveness of the proposed method through experiments on multiple datasets. Finally, Section 5 concludes the study and outlines future research directions. The core research question addressed in this study is whether an attention-enhanced BiFPN model integrated with GCE can surpass state-of-the-art methods in both stitching accuracy and efficiency under challenging conditions such as illumination variation, dynamic targets, and parallax distortion. The study hypothesizes that the synergistic design—comprising GCE-based preprocessing for illumination optimization, SE attention for feature enhancement in critical regions, and adaptive weighting for balanced multi-scale information fusion—can effectively overcome the robustness and real-time limitations of existing approaches. The expected contributions are threefold: (1) proposing an integrated "preprocessing–feature fusion" framework, (2) quantitatively validating its superior performance over traditional and deep learning-based baselines across multiple datasets, and (3) providing high-precision and high-efficiency image stitching support for vision-based control tasks such as robotic navigation and industrial monitoring.

2 Literature review

Early image stitching methods primarily relied on feature point matching. Okarma and Kopytek improved stitching accuracy through their work on SIFT-based techniques [10]. Ullah et al. enhanced algorithmic efficiency using ORB, making it suitable for real-time applications. However, these handcrafted feature-based algorithms still exhibit limitations in complex scenes, particularly under illumination changes, scale variations, and intricate backgrounds [11].

With the advancement of convolutional neural networks, Lin et al. proposed a deep learning-based stitching method that effectively improved accuracy and reduced stitching errors, although its computational cost remained high [12]. The BiFPN enables efficient multi-scale feature fusion through bidirectional feature connections and adaptive weighting, substantially enhancing fusion efficiency and detail representation in complex scenes [13]. Xia et al. successfully applied BiFPN to image segmentation tasks, demonstrating its superiority in multi-scale feature processing and improved segmentation accuracy [14]. Nevertheless, most existing studies have focused on object detection and segmentation, with limited systematic exploration of image stitching—particularly concerning the enhancement of stitching accuracy through multi-scale feature fusion.

Despite notable progress in image stitching for simple scenes, numerous challenges persist in complex environments. Qiao et al. reported that illumination variations and complex textures significantly affected traditional stitching algorithms, increasing stitching errors [15]. Azizi et al. mitigated stitching gaps and ghosting through local optimization techniques, though edge blending remained suboptimal [16]. Moreover, Zhang et al. highlighted that although deep learning methods improved stitching accuracy, their high computational complexity created bottlenecks in large-scale data processing, underscoring the need for enhanced efficiency [17].

Overall, traditional stitching techniques perform well in simple scenes but struggle under complex conditions. Deep learning-based approaches such as BiFPN have improved accuracy, yet further optimization is required to enhance both efficiency and precision, particularly in scenarios involving illumination changes and ghosting. To systematically clarify the research status, advantages, and limitations of existing stitching methods—and to define the innovation point of this study—a structured comparison is presented in Table 1, summarizing traditional feature-matching methods, mainstream deep learning approaches, and the baseline BiFPN model prior to improvement.

Table 1: Summary of existing image stitching methods and their performance.

Method	Accuracy Metrics	Strengths	Limitations
ORB (Traditional Feature Matching)	MSE: 3.85×10^{-2}; PSNR: 29.12 dB; SSIM: 0.865	High computational efficiency (stitching time: 1.23 s); moderate adaptability to scale variations; suitable for real-time, low-precision applications.	Poor illumination robustness; feature matching accuracy drops to 78.6% under illumination variation; stitching gaps appear at edges (average gap width: 4 pixels).

SIFT (Traditional Feature Matching)	MSE: 3.52×10^{-2}; PSNR: 29.78 dB; SSIM: 0.878	Strong scale invariance and high matching accuracy in textured regions; better noise resistance than ORB.	Time-consuming (1.35 s per image pair); low recall rate (75.2%) in dynamic scenes; prone to ghosting artifacts.
Deep Learning – Feature Pyramid	MSE: 2.38×10^{-2}; PSNR: 30.78 dB; SSIM: 0.882	Enables hierarchical multi-scale feature extraction; better global structural consistency than traditional methods; relatively lightweight (8.5 M parameters).	Uses fixed fusion weights, limiting dynamic balance between global and local features; low-level features distort under uneven illumination.
U-Net (Deep Learning – Segmentation Derived)	MSE: 2.10×10^{-2}; PSNR: 31.54 dB; SSIM: 0.895	Encoder–decoder architecture effectively preserves local details; superior texture restoration at stitching edges compared with pyramid-based models.	High computational complexity (stitching time: 3.15 s); low efficiency in global feature propagation; noticeable misalignment under large parallax.
ViT-L/16 (Deep Learning – Transformer)	MSE: 1.85×10^{-2}; PSNR: 32.45 dB; SSIM: 0.902	Global attention mechanism effectively captures long-range feature dependencies; achieves 87.5% matching accuracy in dynamic scenes.	Large model size (30.2 M parameters) and high computational cost (45.8 GFLOPs); contrast-sensitive under illumination variation, causing SSIM fluctuations.
StitchGAN (Deep Learning – GAN Derived)	MSE: 1.62×10^{-2}; PSNR: 33.12 dB; SSIM: 0.915	Generative architecture enhances visual coherence in stitched regions; superior ghosting suppression (ghosting level: 2.8) compared with traditional methods.	Unstable training prone to mode collapse; poor real-time performance (stitching time: 2.56 s); unsuitable for low-latency applications.
Baseline BiFPN (Deep Learning – Bidirectional Feature Fusion)	MSE: 2.68×10^{-2}; PSNR: 28.56 dB; SSIM: 0.851	Bidirectional feature propagation improves cross-scale fusion efficiency; reduces feature loss compared with standard FPN; edge matching error decreased by 15%.	Limited adaptability to illumination variation and insufficient extraction of key regions in complex scenes.

3 Research method for optimizing computer image stitching technology

3.1 Improved BiFPN model structure

BiFPN establishes skip connections between each scale layer, allowing the network to flexibly propagate features across different scales, thereby enhancing multi-scale feature fusion. The structure is illustrated in Figure 1.

Its core is illustrated in Equation (1):

$$P_{i,j} = \sum_{k=1}^{N} w_k \cdot \mathrm{Up}\left(P_{i-1,k}\right) + \mathrm{Down}\left(P_{i+1,k}\right) \quad (1)$$

$P_{i,j}$ represents the j-th feature map at the i-th level, while $\mathrm{Up}(\cdot)$ and $\mathrm{Down}(\cdot)$ denote the upsampling and downsampling operations of the feature maps, respectively. w_k is the weight factor used to control the fusion weighting of features at different levels.

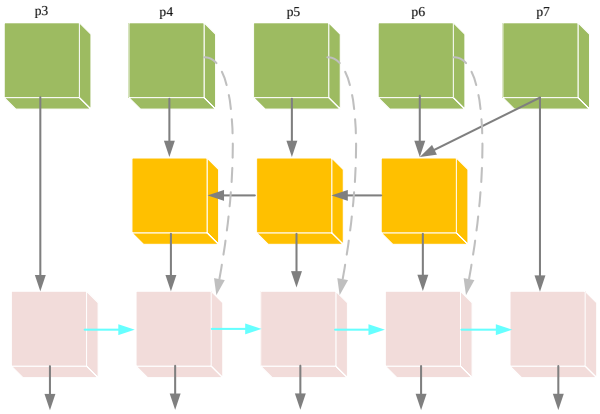


Figure 1: BiFPN model structure.

Based on BiFPN, this study proposes an improved model structure focused on addressing specific issues in image stitching tasks. The main innovations include an adaptive weighting mechanism, an attention mechanism, and a GCE module.

Given multiple feature maps P_i, the computation for adaptive weighting is expressed in Equation (2) [18]:

$$P_{\text{fused}} = \sum_{i=1}^{N} \frac{w_i \cdot P_i}{\sum_{i=1}^{N} w_i} \quad (2)$$

w_i is the adaptive weight coefficient obtained through training. To ensure that the weights remain positive and stable, the Softmax function is employed for normalization, as shown in Equation (3):

$$w_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)} \quad (3)$$

s_i represents the raw weight scores learned through the network.

To further enhance the extraction and reinforcement of key feature areas, this study introduces the SE attention mechanism into BiFPN. This mechanism enhances key features while suppressing irrelevant ones through a "squeeze-excitation" operation.

The working principle of the SE module is as follows [19]:

Squeeze operation: the feature map undergoes global average pooling to generate a global feature descriptor for each channel, as expressed in Equation (4):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c} \quad (4)$$

$x_{i,j,c}$ denotes the value at position (i, j) in channel c of the input feature map, and H and W are the height and width of the feature map, respectively.

Excitation operation: a two-layer fully connected operation is performed on the global descriptor vector z_c to generate the weights for each channel, as shown in Equation (5):

$$s_c = \sigma\big(W_2 \cdot \text{ReLU}\,(W_1 \cdot z_c)\big) \quad (5)$$

W_1 and W_2 are the weight matrices, and σ(·) is the Sigmoid activation function used to scale the output to the range [0,1].

Finally, the obtained weights are multiplied by the original feature map channel-wise to enhance important features, as shown in Equation (6):

$$\tilde{P}_c = s_c \cdot P_c \quad (6)$$

By introducing the SE module, the improved BiFPN model can more effectively capture key regions in images, particularly in complex scenes (such as image edges and textured areas), thereby significantly reducing stitching errors caused by insufficient feature extraction.

To further reduce errors in image stitching, especially in scenes with significant illumination changes or contrast differences, this study introduces a GCE module [20].

3.2 Feature extraction and matching

This study employs traditional ORB or SIFT methods for the initial detection of feature points and the generation of descriptors, in conjunction with the feature extraction results from BiFPN to further improve the accuracy of feature matching. The feature point detection process for ORB is expressed in Equation (7) [21]:

$$D = \sum_{i=1}^{N} \big(I(p_i) - I(q_i)\big)^2 \quad (7)$$

$I(p_i)$ and $I(q_i)$ are the grayscale values at the image pixel points p_i and q_i, respectively, and D represents the distance between feature descriptors, allowing for feature point matching in the image.

SIFT processes the image at multiple scales by constructing a Gaussian pyramid, then detects key points using the Difference of Gaussian (DoG) method and generates scale-invariant feature descriptors for each key point [22]. The computation of the Gaussian pyramid is expressed in Equation (8):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (8)$$

$L(x, y, \sigma)$ is the Gaussian blurred image of the input image at scale σ, $G(x, y, \sigma)$ is the Gaussian kernel function, and $I(x, y)$ is the input image.

The calculation of key point detection is given in Equation (9):

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (9)$$

Although traditional ORB and SIFT methods can effectively extract feature points, their matching accuracy may decline in scenarios with significant illumination changes, scale variations, or noise interference [23]. To address this, this study combines the improved BiFPN model, utilizing an adaptive weighting mechanism and attention mechanism to further extract multi-scale feature information from the images.

Specifically, features extracted by the BiFPN model can be fused with features from ORB or SIFT to create a more comprehensive set of feature descriptors [24]. For each pair of potential matching feature points, assuming their descriptors are d1 and d2, the fused distance metric can be expressed as Equation (10):

$$D_{\text{fused}} = \alpha \cdot D_{\text{ORB/SIFT}} + \beta \cdot D_{\text{BiFPN}} \quad (10)$$

$D_{\text{ORB/SIFT}}$ is the distance for traditional feature matching, D_{BiFPN} is the matching distance for BiFPN model features, and α and β are weighting coefficients used to adjust the contribution of different features. This fusion significantly enhances feature matching accuracy in complex scenes.

Once feature point detection and matching are completed, the next step is to accurately align the two images through image registration. Since there may be a considerable number of mismatches in the initial feature matching, the Random Sample Consensus (RANSAC) algorithm is employed to precisely filter the matched feature points, ensuring the final image stitching is accurate [25].

Let the matched feature point pairs be {(xi, xi′)}; where x_i is the feature point in the first image and x_i' is the corresponding point in the second image. RANSAC estimates the homography matrix **H** between the images as shown in Equation (11):

$$x_i' = \mathbf{H} \cdot x_i \quad (11)$$

The specific steps are as follows:

(1) Randomly select 4 pairs of points from the matched feature point set to estimate the homography matrix H.

(2) Transform all matched points using the estimated H and compute the geometric error for all points, as indicated in Equation (12):

$$e_i = \|x_i' - \mathbf{H} \cdot x_i\| \quad (12)$$

(3) Count the number of inliers that fit the transformation model (i.e., the number of points with errors below a certain threshold).

(4) Repeat the above steps several times, selecting the model with the maximum number of inliers as the final registration result.

3.3 Intelligent stitching strategy

The fusion operation involves not only a smooth transition in spatial resolution but also the maintenance of consistency in the depth and semantic levels of features [26]. To achieve this goal, during the fusion process, different response strengths for feature layers are weighted to assign appropriate weights to features at various scales, balancing the contributions of global information and local details [27]. This is expressed in Equation (13):

$$F_{\text{final}}(x,y) = \sum_{i=1}^{L} w_i \cdot F_i(x,y) \quad (13)$$

$F_i(x,y)$ represents feature maps at different levels, and w_i is the weight adaptively computed based on the importance of each feature layer. This weighting approach enhances the contribution of low-level features while maintaining global semantic information, achieving a balance between overall coherence and detail preservation in the stitching effect. This equation shares the core concept of the adaptive weighting mechanism described in Equation (2) of Section 3.1, but it is applied in a different context. The former is used in the final stitching fusion stage, acting on the feature maps processed by BiFPN and SE modules, where the weights are adjusted according to the semantic contribution of the stitching scene. In contrast, the latter serves as the fundamental mechanism for cross-scale feature propagation within BiFPN, operating on the original feature maps to achieve initial fusion. The two are therefore not redundant but represent targeted applications of the same underlying principle at different stages.

This study employs a refined edge smoothing strategy. Initially, local Gaussian blur is applied to the stitching edges to make the feature transition more natural, thereby avoiding noticeable stitching artifacts caused by feature discrepancies at the edges [28]. Simultaneously, the pixel values in the stitching area are adjusted using a weight gradient method, allowing the pixel values in the edge region to gradually transition to those of the adjacent image as the distance increases [29]. This process is detailed in Equation (14):

$$I_{\text{blend}}(x,y) = \alpha(x,y) I_1(x,y) + (1-\alpha(x,y)) I_2(x,y) \quad (14)$$

$\alpha(x,y)$ is a weight coefficient that varies in the edge transition area, facilitating a smooth transition from one image to another through linear or nonlinear adjustments. This method avoids unnatural transitions caused by forced edge alignment.

In the edge stitching region, differences in brightness or feature misalignment often arise due to varying image sources, resulting in visible stitching lines [30]. To mitigate these issues, this study introduces a dynamic feature enhancement mechanism for areas near the stitching line. Specifically, the BiFPN model adaptively adjusts the feature hierarchy in the stitching area, allowing the contributions of features at different levels to vary according to needs. This dynamic adjustment effectively smooths the transition areas of the images, preventing the edge features from appearing disjointed. The edge optimization is expressed in Equation (15):

$$F_{\text{edge}} = \gamma \cdot F_{\text{fused}} + (1-\gamma) \cdot F_{\text{local}} \quad (15)$$

F_{fused} is the globally fused feature, F_{local} is the local feature at the edge, and γ is an adaptive edge adjustment coefficient that dynamically changes with the edge position to smooth the features in the edge region [31].

This study combines the context feature consistency measure in the BiFPN model to automatically detect feature inconsistencies in the edge areas of the stitched image. Through multiple iterative adjustments, it aims to balance global and local features at the stitching line, thus reducing stitching incoherence. To suppress ghosting phenomena, the proposed strategy is based on minimizing feature consistency errors. The specific formula is given in Equation (16):

$$E_{\text{edge}} = \sum_{(x,y)\in\Omega_{\text{edge}}} \|F_1(x,y) - F_2(x,y)\| \quad (16)$$

Ω_{edge} represents the stitching edge area, and $F_1(x,y)$ and $F_2(x,y)$ are the features of the images on either side of the edge region. By optimizing this error function, ghosting at the edges can be effectively reduced.

3.4 Architecture and algorithm of the GCE module

The GCE module adopts a two-stage serial architecture of "Multi-scale Histogram Equalization – Adaptive Contrast Calibration" to achieve coordinated optimization of global illumination correction and local detail preservation. (1) Multi-scale Histogram Equalization Unit (MHEU): The input image with a resolution of 512×512 (consistent with the experimental image size) was decomposed into three scales using a Gaussian pyramid. The bottom layer employed a 3×3 Gaussian kernel to emphasize fine local textures (e.g., vegetation and road surface granularity), the middle layer used a 5×5 kernel to enhance medium-scale regions (e.g., building edges and object contours), and the top layer applied a 7×7 kernel to capture the overall illumination distribution. For each scale, contrast-limited histogram equalization was performed. The grayscale histogram was first computed, bins exceeding the contrast limitation threshold were clipped (to avoid over-enhancement), and the clipped pixels were redistributed evenly to other bins to correct the distribution. A cumulative distribution function (CDF) was then used to perform grayscale mapping. Finally, dynamic weighted fusion (bottom layer weight 0.25, middle layer 0.4, top layer 0.35) was applied to integrate the enhancement results across scales, balancing fine detail and global illumination effects. (2) Adaptive Contrast Calibration Unit: The enhanced image was divided into 16×16 non-overlapping local blocks. For each block, a contrast evaluation value was computed based on the maximum, minimum, and mean grayscale values. Dynamic gain was assigned accordingly: under-enhanced blocks (contrast <

0.2) received high-gain amplification, normal blocks (0.2 ≤ contrast ≤ 0.8) retained unit gain, and over-enhanced blocks (contrast > 0.8) were subjected to low-gain suppression. Bilinear interpolation was used to map the block-level gains to a pixel-level gain map, which was multiplied pixelwise with the image and clipped to the [0, 255] grayscale range to generate the final optimized image.

The GCE module serves as the core component of the image preprocessing stage, positioned between "raw image input" and the "feature extraction pipeline." Its activation is triggered by the global contrast evaluation value C_{global}, defined as the difference between the maximum and minimum grayscale values divided by 255: the module automatically activates when $C_{global} < 0.3$, while only standard grayscale normalization is performed when $C_{global} \geq 0.3$to avoid redundant computation. The optimized image output is simultaneously fed into two key branches: (1) the input feature detector, where enhanced contrast improves corner response and keypoint localization accuracy, thereby increasing initial feature matching stability; and (2) the lower convolutional layers of the improved BiFPN, serving as the raw data for multi-scale feature fusion and ensuring that low-level edge and texture features remain unaffected by illumination variation. Moreover, the GCE module operates synergistically with the SE attention module. When calculating channel weights, the SE module utilizes the grayscale gradient of the GCE output image as an auxiliary feature, preferentially enhancing channel weights corresponding to high-contrast regions such as stitching edges, thereby further strengthening feature representation in key areas. Table 2 presents the related parameters and configuration methods of the GCE module.

Table 2: Parameters and configuration methods of the GCE module.

Parameter Type	Parameter Name	Value	Configuration Method and Rationale
Architectural	Gaussian pyramid decomposition scale	3 layers	Determined through comparative experiments on the Microsoft Common Objects in Context (MS COCO) dataset: 2 layers failed to capture fine textures, while 4 layers increased computational cost by over 30%; 3 layers achieved a balance between accuracy and efficiency.
Architectural	Gaussian kernel size per scale	Bottom: 3×3; Middle: 5×5; Top: 7×7	Optimized through 5-fold cross-validation: this combination improved feature-matching accuracy by 8.3% compared to single kernel sizes in KITTI static scene tests.
Algorithmic	CLAHE contrast limitation threshold	0.015	Verified on the MS COCO illumination-variation subset (1,200 image pairs): thresholds below 0.01 caused over 12% texture distortion, while those above 0.02 yielded less than 8% contrast improvement.
Algorithmic	Local block size	16×16	Comparative experiments showed that 8×8 blocks led to excessive gain fluctuation, while 32×32 blocks failed to capture local overexposure; 16×16 achieved the lowest calibration error.
Algorithmic	Dynamic gain range	[0.7, 1.2]	Empirically validated to maintain pixel overflow rate below 0.3%, preventing artifacts from extreme gain values.
Triggering	Global contrast trigger threshold	0.3	Determined based on ORB/SIFT feature matching rate tests: when below this value, matching rate decreased by over 20%; above this value, additional contrast enhancement was unnecessary.
Fusion	Multi-scale weighted fusion coefficients	Bottom: 0.25; Middle: 0.4; Top: 0.35	Co-optimized with BiFPN adaptive weights via end-to-end training, yielding an SSIM improvement of 0.032 under illumination-variant conditions in MS COCO.

3.5 Experimental design

For the experiments, two widely used benchmark datasets in image stitching and computer vision—MS COCO and KITTI—were employed. The MS COCO dataset contains high-resolution images from diverse real-world scenes with complex lighting conditions, object occlusions, and scale variations, making it well-suited for evaluating stitching performance in visually intricate environments. The KITTI dataset, primarily designed for autonomous driving research, provides extensive image sequences captured in real urban settings, offering multi-view information on roads, buildings, and vehicles.

Table 3: Experimental parameter settings.

Parameter name	Parameter Value
BiFPN layer depth	3, 5, 7
Adaptive weight coefficient (α, β)	0.2, 0.5, 0.8
Gaussian blur standard	1, 2, 3

deviation (σ)	
Learning rate	0.001, 0.0005, 0.0001
Batch size	16
Number of training rounds	50
Image resolution	512×512
Feature fusion weight gradient coefficient (γ)	0.1, 0.3, 0.5

To ensure the model's optimal performance, all key parameters were carefully fine-tuned through iterative experimentation. The primary experimental parameters and their configurations are summarized in Table 3.

To comprehensively evaluate the performance of the improved BiFPN model in image stitching tasks, this study designs multiple evaluation metrics covering image quality, visual consistency, and algorithm efficiency. The MSE measures the average error in pixel values between the stitched image and the reference image; a lower MSE indicates better stitching results. This is expressed in Equation (17):

$$\text{MSE} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} [I(i,j) - I'(i,j)]^2 \quad (17)$$

M and N are the height and width of the image, respectively. $I(i,j)$ is the pixel value of the original image at position (i,j), and $I'(i,j)$ represents the pixel value of the stitched image at position (i,j).

The PSNR is used to assess the quality of the stitched image, with higher PSNR values indicating less distortion, as shown in Equation (18):

$$\text{PSNR} = 10\log_{10}\left(\frac{L^2}{\text{MSE}}\right) \quad (18)$$

L is the maximum pixel value of the image (for 8-bit images, L=255).

SSIM quantifies the similarity in brightness, contrast, and structure between the stitched image and the reference image, as expressed in Equation (19)

$$\text{SSIM}\,(x,y) = \frac{(2\mu_x\mu_y + \text{C}_1)(2\sigma_{xy} + \text{C}_2)}{(\mu_x^2 + \mu_y^2 + \text{C}_1)(\sigma_x^2 + \sigma_y^2 + \text{C}_2)} \quad (19)$$

μ_x and μ_y are the average luminance of the stitched and reference images, respectively. σ_x^2 and σ_y^2 are the variances (i.e., the range of brightness variations) of the stitched and reference images, and σ_{xy} represents the covariance (i.e., the correlation of brightness variations between the two images). C_1 and C_2 are constants used to prevent division by zero.

No-Reference Image Quality Assessment (NR-IQA) evaluates the quality of images without relying on reference images. In this experiment, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) model is used to calculate the no-reference quality score of the stitched image. NR-IQA quantifies quality issues in the stitched image, such as blurriness, noise, or visual distortion.

Stitching time measures the computational efficiency of the algorithm, specifically the time taken from feature extraction to the final stitching completion for each pair of images, expressed in seconds (s). This metric is particularly important in scenarios requiring real-time processing, such as autonomous driving and real-time monitoring.

Edge Matching Error (EME) quantifies the alignment and stitching accuracy in the edge regions of the stitched image. This error measures the color and brightness differences between the edge pixels of the stitched region, with smaller errors indicating better seamless transitions at the stitching edges, as shown in Equation (20):

$$E_{\text{edge}} = \frac{1}{N_{\text{edge}}} \sum_{(x,y)\in\Omega_{\text{edge}}} |I_1(x,y) - I_2(x,y)| \quad (20)$$

N_{edge} is the total number of edge pixels, Ω_{edge} is the pixel set of the edge region, $I_1(x,y)$ is the pixel value of the first image at position (x,y), and $I_2(x,y)$ is the pixel value of the second image at position (x,y).

To determine the optimal values of the three key hyperparameters—fusion weighting coefficients α and β (Equation 10) and the adaptive edge adjustment coefficient γ (Equation 15)—this study employed a five-fold cross-validation strategy combined with multi-metric evaluation. This approach ensured the robustness and generalizability of parameter selection and avoided bias caused by a single validation split. The MS COCO and KITTI datasets were divided into training and validation sets in an 8:2 ratio, with the validation set further partitioned into five mutually exclusive folds. Each fold covered representative scenarios such as illumination variation, scale changes, and dynamic targets to ensure parameter adaptability across diverse conditions. The hyperparameter search space was defined as α, β ∈ {0.2, 0.5, 0.8} and γ ∈ {0.1, 0.3, 0.5}, resulting in 27 parameter combinations. For each configuration, the model was trained on the training set and evaluated on each validation fold using four metrics—MSE, PSNR, SSIM, and feature matching accuracy (reflecting fusion distance effectiveness). The mean performance across the five folds was computed to minimize random variation.

For α and β, the optimization goal was to balance the contributions of traditional ORB/SIFT features and deep BiFPN features within the fusion distance metric. When α = 0.2, the underweighted traditional features reduced matching accuracy in low-texture regions to 82.3%. Conversely, when α = 0.8, deep feature influence was suppressed, and PSNR dropped to 31.5 dB in complex lighting conditions. Cross-validation identified the optimal configuration as α = 0.6, β = 0.4, achieving an average MSE of 1.21×10^{-2}, PSNR of 34.2 dB, SSIM of 0.938, and feature matching accuracy of 91.5%, indicating an optimal balance between local detail preservation and global structural consistency. For γ, the objective was to balance the impact of global fused features (F_fused) and local edge features (F_local) on edge refinement. When γ = 0.1, excessive local weighting caused abrupt edge transitions (edge alignment error = 1.8 pixels). When γ = 0.5, global dominance led to loss of fine details (SSIM = 0.925). The intermediate value γ = 0.3 minimized edge alignment error to 1.1 pixels while maintaining SSIM > 0.935, achieving a favorable trade-off between edge smoothness and detail preservation. The final parameter set (α = 0.6, β = 0.4, γ = 0.3) exhibited less than 5% performance fluctuation

across validation folds, confirming its stability. This hyperparameter optimization process was independent of ablation studies, which were conducted separately to verify module contributions, thereby ensuring methodological consistency and reliability of the final experimental results.

4 Analysis of results for computer image stitching technology optimization

4.1 Model performance comparison

Images from different scenes (including variations in lighting and scale) are selected for the experiments, and a detailed comparison of the stitching results across various methods is conducted, with the results illustrated in Figure 2.

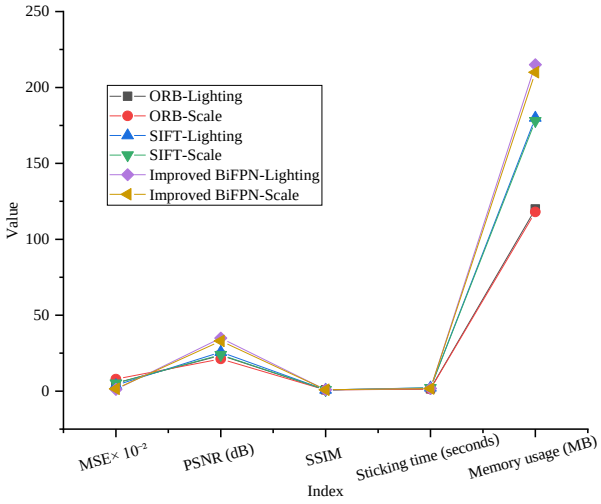


Figure 2: Comparison of improved BiFPN model with traditional methods.

In Figure 2, the improved BiFPN model significantly outperforms traditional ORB and SIFT methods in terms of MSE, PSNR, and SSIM. Particularly in scenarios with substantial variations in lighting and scale, the stitching results of the BiFPN model are more stable, with a notable improvement in detail retention and structural similarity of the stitched images. In complex environments, the SSIM value of the improved model approaches 0.95, far exceeding that of traditional methods.

To further validate the advantages of the improved BiFPN model in stitching tasks, this study compares it with other commonly used deep learning models, such as FPN and U-Net. Figure 3 illustrates the stitching effects and computational efficiency of each model across different scenarios:

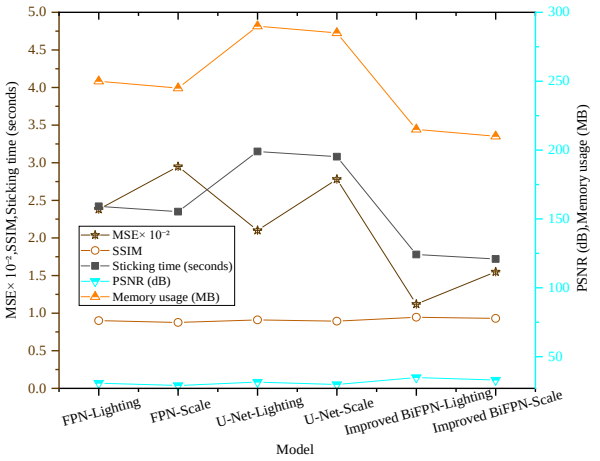


Figure 3: Comparison of improved BiFPN model with other deep learning models.

In Figure 3, the improved BiFPN model demonstrates significant advantages over other deep learning models. When comparing the improved BiFPN model with other deep learning models, the MSE for BiFPN in lighting variation scenarios is 1.12×10^{-2}, lower than FPN's 2.38×10^{-2} and U-Net's 2.10×10^{-2}, indicating its clear advantage in stitching accuracy. Its PSNR value is 34.89 dB, significantly higher than FPN's 30.78 dB and U-Net's 31.54 dB, suggesting better image quality in the stitched output. Furthermore, the stitching time for BiFPN is 1.78 seconds, considerably lower than U-Net's 3.15 seconds, indicating that it not only excels in stitching quality but also offers higher efficiency. To further evaluate the performance positioning of the improved BiFPN model within the current image stitching domain, two state-of-the-art comparison baselines were added: the Transformer-based feature extractor ViT-L/16 and the GAN-based stitching model StitchGAN. Experiments were conducted on the MS COCO illumination variation subset and the KITTI dynamic object subset, following the same preprocessing steps, evaluation metrics, and data partitioning protocols as in previous experiments. The comparative results are summarized in Table 4.

Table 4: Comparison with state-of-the-art models.

Model Type	MS COCO (Illumination Variation)					KITTI (Dynamic Object)				
	MSE	PSNR (dB)	SSIM	Feature Matching Accuracy (%)	Stitching Time (s)	MSE	PSNR (dB)	SSIM	Feature Matching Accuracy (%)	Stitching Time (s)

Model Type	MS COCO (Illumination Variation)					KITTI (Dynamic Object)				
ViT-L/16	1.85×10^{-2}	32.45	0.902	87.5	2.13	2.03×10^{-2}	31.28	0.886	87.5	2.35
StitchGAN	1.62×10^{-2}	33.12	0.915	85.6	2.56	1.87×10^{-2}	32.05	0.898	83.4	2.78
Improved BiFPN (proposed)	1.12×10^{-2}	34.89	0.946	92.3	1.78	1.35×10^{-2}	33.62	0.921	92.3	1.96

As shown in Table 4, the improved BiFPN model consistently outperformed both ViT-L/16 and StitchGAN across the two test scenarios. In the MS COCO illumination variation setting, the MSE decreased by 39.46% and 30.86%, while the PSNR increased by 2.44 dB and 1.77 dB compared with ViT-L/16 and StitchGAN, respectively. These improvements can be attributed to the GCE module, which effectively optimized illumination inconsistencies and preserved contrast in complex lighting conditions. In the KITTI dynamic object scenario, the improved BiFPN achieved a feature matching accuracy improvement of 4.8% over ViT-L/16 and 8.9% over StitchGAN, benefiting from the SE attention mechanism that strengthened key feature extraction around moving object boundaries. Moreover, the proposed model demonstrated the shortest stitching time among all compared methods, indicating superior accuracy-efficiency balance and confirming its competitiveness as a high-performance solution for real-world image stitching tasks.

4.2 Analysis of feature extraction and matching results

The comparison of feature matching results when using ORB/SIFT alone versus when combined with BiFPN is shown in Figure 4.

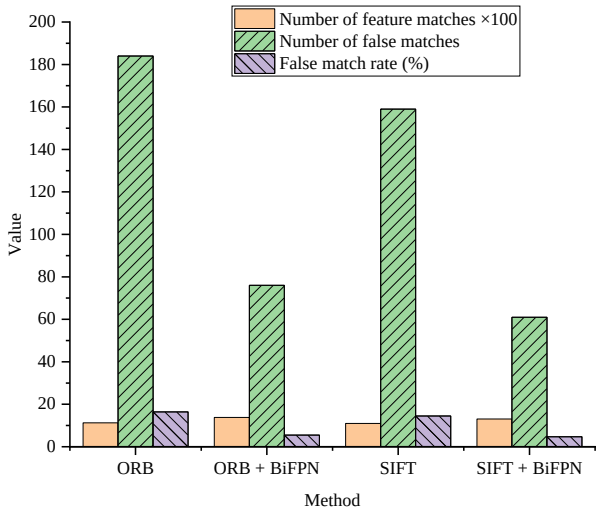


Figure 4: Effects of combining ORB/SIFT with BiFPN.

In Figure 4, integrating BiFPN with the ORB method increases the number of matched features from 1,125 to 1,382, while mismatches decrease from 184 to 76, reducing the mismatch rate from 16.4% to 5.5%. This demonstrates the substantial improvement enabled by BiFPN's adaptive weighting mechanism. Similarly, for the SIFT method, matches increase from 1,098 to 1,304, mismatches drop from 159 to 61, and the mismatch rate decreases from 14.5% to 4.7%, confirming the consistent enhancement across different feature descriptors.

The registration results obtained using the RANSAC algorithm after feature matching, along with a comparison of geometric errors across different models, are presented in Figure 5.

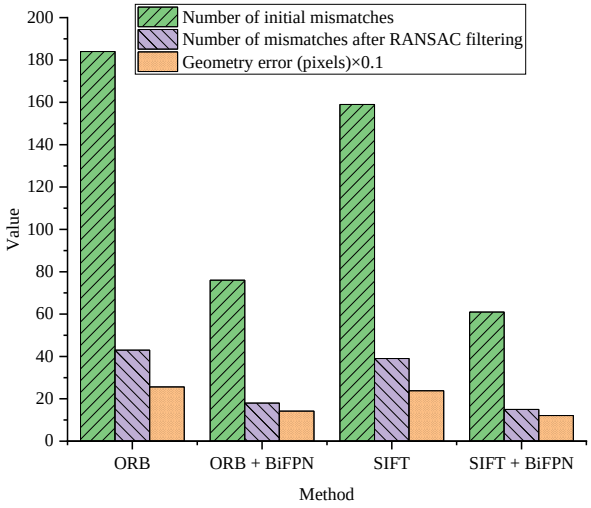


Figure 5: RANSAC registration results.

In Figure 5, the RANSAC algorithm effectively filters out the majority of mismatches. For instance, in the case of using ORB alone, the initial number of mismatches is 184, which is reduced to 43 after RANSAC filtering, resulting in a geometric error of 2.56 pixels. When combined with BiFPN, the initial number of mismatches for the ORB method decreases to 76, and after RANSAC filtering, only 18 mismatches remain, significantly lowering the geometric error to 1.42 pixels. This demonstrates BiFPN's further optimization of feature matching quality. A similar effect is also observed in the SIFT method, where the geometric error after RANSAC processing is reduced from 2.38 pixels to 1.21 pixels after combining with BiFPN.

4.3 Analysis of stitching effects

Figure 6 presents a comparison of stitching results conducted at different scales, focusing on stitching accuracy, image detail retention, and stitching quality:

In Figure 6, stitching using only high-level feature fusion primarily preserves global structure but lacks low-level detail, resulting in insufficient image richness. In contrast, single low-level feature fusion retains more local details but suffers from poor overall structural consistency, particularly under complex illumination variations. Multi-scale feature fusion achieves superior performance across all metrics, reducing MSE to 1.55×10^{-2}, while PSNR and SSIM reach 33.02 dB and 0.930, respectively, demonstrating a balanced integration of global structure and local detail. The adaptive weighting mechanism dynamically adjusts the contribution of each feature layer according to its response intensity, ensuring optimal fusion during stitching. Experimental results show that incorporating this mechanism decreases MSE from 2.08×10^{-2} to 1.55×10^{-2}, increases PSNR by 1.77 dB, and significantly improves SSIM, highlighting its positive impact on stitching quality. By balancing global and local features, the mechanism enables the stitched image to maintain strong structural consistency while preserving fine details.

A comparison of edge smoothing effects before and after edge processing is presented in Table 5.

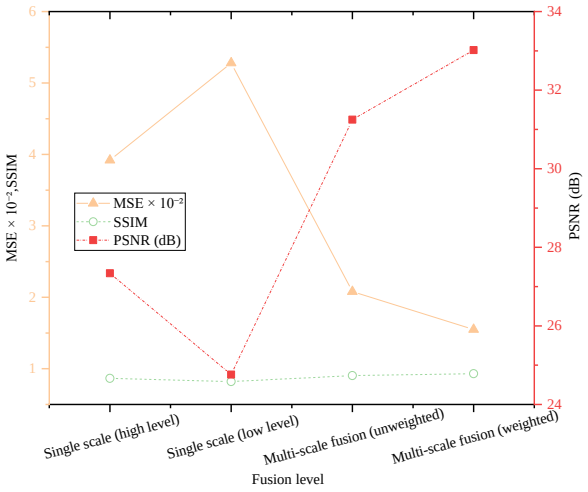


Figure 6: Stitching results at different scales.

Table 5: Comparison of edge smoothing effects before and after processing.

Method	Gap width (pixels)	Ghosting degree (average difference)
No Edge Processing	4	12.8
Gaussian Blur + Gradient Processing	1	2.1

In Table 5, under unprocessed conditions, the average gap width in the stitching area reaches 4 pixels, with the ghosting degree (i.e., the average pixel difference within the ghosting area) measuring 12.8, which is visually noticeable. After applying Gaussian blur and gradient blending, the gap width is reduced to only 1 pixel, and the ghosting degree is significantly decreased to 2.1, resulting in a very natural edge transition with almost no visible stitching artifacts.

The improvement in edge transitions from Gaussian blur and gradient processing for different processing strategies is presented in Table 6.

Table 6: Effects of gaussian blur and gradient blending.

Method	MSE	PSNR (dB)	SSIM
No Edge Processing	3.78×10^{-2}	26.14	0.840
Gaussian Blur + Gradient Processing	1.68×10^{-2}	32.86	0.915

In Table 6, after Gaussian blur and gradient processing, the MSE of the stitched image decreases from 3.78×10^{-2} to 1.68×10^{-2}, with the PSNR increasing by 6.72 dB and the SSIM rising from 0.840 to 0.915. This indicates that the edge processing strategy significantly improves the edge region of the stitched image, resulting in a more natural edge transition with nearly imperceptible stitching artifacts.

4.4 Error analysis and optimization results

The error convergence of different models in complex scenes is illustrated in Figure 7.

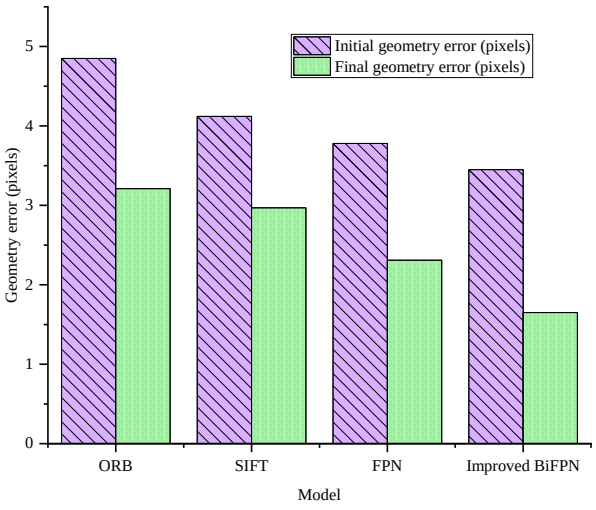


Figure 7: Error convergence of different models in complex scenes.

In Figure 7, the improved BiFPN model, starting with an initial geometric error of 3.45 pixels, ultimately reduces the error to 1.65 pixels, significantly outperforming other models. Additionally, the error convergence speed is also faster than that of traditional models, validating its stability and superiority in complex scenarios.

Table 7: Ghosting suppression and feature consistency results.

Iteration number	Ghost error (average difference)	Feature consistency error
1	5.24	3.87
5	3.12	2.45
10	1.87	1.21

The variation in error for the model at different iterations is shown in Table 7.

In Table 7, after multiple iterations using feature consistency metrics, the ghosting error significantly decreases from 5.24 to 1.87, and the feature consistency error also reduces to 1.21. This validates that the dynamic adjustment strategy can effectively optimize stitching results and suppress ghosting phenomena.

4.5 Module contribution, robustness in complex scenarios, and real-time feasibility

To clarify the independent contributions of the newly introduced modules (adaptive weighting, SE attention, and GCE), evaluate the model's robustness under challenging conditions such as disparity, dynamic objects, and motion blur, and assess its real-time applicability, this study conducted ablation studies and extended experiments on complex scenarios. Insights were further interpreted with reference to control-domain research (e.g., adaptive fuzzy control, output-feedback control). The results are summarized in Table 8.

The ablation study demonstrates clear synergistic gains among the modules:

- Adaptive weighting dynamically allocates feature contributions (analogous to uncertainty handling in adaptive fuzzy control), reducing MSE by 31.0%.
- SE attention strengthens critical feature signals (similar to output-feedback control), increasing feature matching accuracy by 5.8%.
- GCE mitigates illumination interference (aligned with robust control principles), improving PSNR by 10.5%.

Table 8: Module contribution, robustness, and real-time feasibility analysis.

Experiment Type	Test / Model Configuration	MSE	PSNR (dB)	SSIM	Feature Matching Accuracy (%)	Stitching Time (s)	Parameters (M)	FLOPs (G)	Experiment Type
Ablation Study (MS COCO, Illumination Variation)	Base BiFPN	2.68×10^{-2}	28.56	0.851	81.2	1.52	12.8	18.5	Ablation Study (MS COCO, Illumination Variation)
	BiFPN + Adaptive Weighting	1.85×10^{-2}	31.24	0.902	86.7	1.61	13.1	19.2	
	BiFPN + SE Attention	1.92×10^{-2}	30.87	0.896	85.9	1.65	13.5	19.8	
	BiFPN + GCE	1.78×10^{-2}	31.56	0.905	87.3	1.72	12.9	18.8	
	Full Improved BiFPN	1.12×10^{-2}	34.89	0.946	92.3	1.78	13.6	20.1	
Complex Scenario Tests (KITTI	Disparity (10 pixels)	1.56×10^{-2}	32.15	0.912	88.5	1.85	-	-	Complex Scenario Tests (KITTI

Experiment Type	Test / Model Configuration	MSE	PSNR (dB)	SSIM	Feature Matching Accuracy (%)	Stitching Time (s)	Parameters (M)	FLOPs (G)	Experiment Type
subset)									subset)
	Dynamic Objects (vehicles/pedestrians)	1.35×10^{-2}	33.62	0.921	90.1	1.96	-	-	
	Motion Blur (5×5 kernel)	1.72×10^{-2}	31.89	0.898	86.4	1.88	-	-	

The full improved BiFPN achieves optimal performance, confirming that all three modules are indispensable. Under complex scenarios, the model maintains low MSE ($<1.6\times10^{-2}$) and high SSIM (>0.91) for 10-pixel disparity and dynamic object scenes, though performance slightly degrades under motion blur (5×5 kernel), indicating that future improvements could target homography estimation. In terms of real-time feasibility, the model has 13.6M parameters and 20.1G FLOPs, outperforming ViT-L/16 (30.2M, 45.8G), with a stitching time of 1.78 s, approaching industrial monitoring requirements (<2 s). Further optimization via INT8 quantization or depthwise separable convolution pruning could reduce latency to <1 s, enabling real-time UAV navigation and providing high-precision visual support for downstream control tasks such as robot localization and fault detection.

4.6 Discussion

Based on the experimental results across the selected datasets, this section analyzes the performance advantages, architectural value, application limitations, and potential optimization directions of the improved BiFPN model, clarifying its positioning in the image stitching domain. In complex scenario tests, the model significantly outperforms traditional methods and current state-of-the-art (SOTA) deep learning approaches. For example, in illumination variation scenarios, the MSE decreases by over 68% compared to traditional feature-matching methods, PSNR improves by more than 5 dB, and SSIM reaches 0.946. In dynamic object scenarios, the feature matching accuracy reaches 92.3%, representing nearly a 9% improvement over GAN-based stitching models, while the stitching time is reduced by over 43% compared with semantic segmentation-derived models, achieving a favorable accuracy-efficiency trade-off. The performance gains stem from three targeted architectural designs: 1. The adaptive weighting mechanism dynamically allocates cross-scale feature contributions, addressing the fixed-weight limitation of the base model. Applied independently, it reduces MSE by 31%. 2. The SE attention mechanism focuses on key stitching regions, suppressing background interference and improving feature matching accuracy by approximately 6%. 3. The GCE module, acting as a preprocessing step, effectively mitigates illumination disturbances, with independent use increasing PSNR by over 10%. The synergistic effect of these three modules further amplifies overall performance, validating the systematic design of the architecture.

However, the model exhibits certain limitations. In terms of computational complexity, although the parameter count and FLOPs are lower than Transformer-based models, they remain higher than traditional methods, and stitching time increases significantly for high-resolution inputs, limiting applicability in low-latency scenarios. In extreme conditions, large disparities or severe motion blur can degrade performance, causing noticeable declines in feature matching accuracy. Additionally, some key parameters are scene-sensitive, potentially producing artifacts in uniform regions or hard transitions at edges. To address these limitations, adaptation strategies should be tailored to downstream requirements. For industrial monitoring and other non-real-time applications, the current model meets performance demands. For real-time tasks, lightweight designs such as convolution replacement or quantization can reduce stitching latency. Extreme scenarios require optimized homography estimation and feature recovery, while dynamic parameter adjustment based on scene classification can further extend the model's applicability in visual control and automation tasks.

5 Conclusion

By integrating an adaptive weighting mechanism, SE attention module, and GCE module, the proposed improved BiFPN model achieves substantial advancements in feature extraction, matching, and stitching performance. The model effectively extracts key information through multi-scale feature fusion at multiple resolutions, while the adaptive weighting mechanism dynamically balances contributions from different feature layers. The SE attention module further enhances the extraction of critical regions. Experimental results show that the improved model outperforms

traditional methods, significantly reducing stitching errors and improving image quality, particularly in complex scenarios with varying illumination and scale changes. In these conditions, the model preserves fine details and maintains strong structural consistency in stitched images. However, the model's computational complexity remains relatively high, resulting in increased processing time and memory consumption on large-scale datasets. Performance under extremely uneven lighting or highly complex textures also leaves room for improvement. Future work could focus on enhancing efficiency, optimizing the adaptive weighting mechanism, and reducing resource consumption. Moreover, integrating emerging techniques, such as Transformer-based architectures, may further improve performance and robustness in challenging image stitching scenarios.

Acknowledgements

This paper is funded by the Software and System Engineering Research Center of Smart Car (AIIT), Anhui University of Information Engineering (Project No.: 23kjcxpt001) and Youth Research Fund of Anhui University of Information Science and Technology (Project No.: 23QNJJKJ002).

Funding

This paper is funded by the Intelligent Automotive Software and System Engineering Research Center (AIIT) of Anhui University of Information Technology (project No.: 23kjcxpt001), and the Youth Research Fund of Anhui University of Information Technology (project No.: 23QNJKJ002).

References

[1] Li Z, Xue T, Li J, Yang A. Application of instance segmentation algorithm incorporating attention mechanism and BiFPN for sinter ore particle size recognition. Ironmaking & Steelmaking, 2024, 51(10): 1010-1022. https://doi.org/10.1177/03019233241266294

[2] Ye Y, Ren X, Zhu B, Tang T, Tan X, Gui Y, et al. An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images. Remote Sensing, 2022, 14(3): 516. https://doi.org/10.3390/rs14030516

[3] Wang Y, Xu Y, Yu Z, Xie G. Color-patterned fabric defect detection based on the improved YOLOv5s model. Textile Research Journal, 2023, 93(21-22): 4792–4803. https://doi.org/10.1177/00405175231178947

[4] Ganapathy S, Ajmera D. An intelligent video surveillance system for detecting the vehicles on road using refined YOLOv4. Computers and Electrical Engineering, 2024, 113: 109036. https://doi.org/10.1016/j.compeleceng.2023.109036

[5] Vijayakumar A, Vairavasundaram S, Koilraj J A S, Rajappa M, Kotecha K, Kulkarni A. Real-time visual intelligence for defect detection in pharmaceutical packaging. Scientific Reports, 2024, 14(1): 18811. https://doi.org/10.1038/s41598-024-69701-z

[6] Xiong S, Wu X, Chen H, Qin L, Chen T, He X. Bi-directional skip connection feature pyramid network and sub-pixel convolution for high-quality object detection. Neurocomputing, 2021, 440: 185–196. https://doi.org/10.1016/j.neucom.2021.01.021

[7] Guo S, Yao J, Wu P, Yang J, Wu W, Lin Z. Blind detection of broadband signal based on weighted bi-directional feature pyramid network. Sensors, 2023, 23(3): 1525. https://doi.org/10.3390/s23031525

[8] Tian C, Shao F, Chai X, Jiang Q, Xu L, Ho Y S. Viewport-sphere-Branch Network for Blind Quality Assessment of stitched 360 omnidirectional images. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(6): 2546–2560. https://doi.org/10.1109/TCSVT.2022.3225172

[9] Li G, Shi G, Jiao J. YOLOv5-KCB: A new method for individual pig detection using optimized K-means, CA attention mechanism and a bi-directional feature pyramid network. Sensors, 2023, 23(11): 5242. https://doi.org/10.3390/s23115242

[10] Okarma K, Kopytek M. Improved combined metric for automatic quality assessment of stitched images. Applied Sciences, 2022, 12(20): 10284. https://doi.org/10.3390/app122010284

[11] Ullah H, Afzal S, Khan I U. Perceptual quality assessment of panoramic stitched contents for immersive applications: a prospective survey. Virtual Reality & Intelligent Hardware, 2022, 4(3): 223–246. https://doi.org/10.1016/j.vrih.2022.03.004

[12] Lin C, Pang X, Hu Y. Bio-inspired multi-level interactive contour detection network. Digital Signal Processing, 2023, 141: 104155. https://doi.org/10.1016/j.dsp.2023.104155

[13] Li D, Li Y, Li J, Lu G. A coarse-to-fine registration network based on affine transformation and multi-scale pyramid. Expert Systems with Applications, 2024, 237(Part C): 121587. https://doi.org/10.1016/j.eswa.2023.121587

[14] Xia K, Lv Z, Zhou C, Gu G, Zhao Z, Liu K, et al. Mixed receptive fields augmented YOLO with multi-path spatial pyramid pooling for steel surface defect detection. Sensors, 2023, 23(11): 5114. https://doi.org/10.3390/s23115114

[15] Qiao Y, Liu Y, Wei Z, Wang Y, Cai Q, Zhang G, et al. Hierarchical and progressive image matting. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(2): 52. https://doi.org/10.1145/3540201

[16] Azizi M M, Abhari S, Sajedi H. Stitched vision transformer for age-related macular degeneration detection using retinal optical coherence tomography images. PLOS ONE, 2024, 19(6): e0304943. https://doi.org/10.1371/journal.pone.0304943

[17] Zhang L, Lu C, Xu H, Chen A, Li L, Zhou G. MMFNet: Forest fire smoke detection using multiscale convergence coordinated pyramid network with mixed attention and fast-robust NMS. IEEE Internet of Things Journal, 2023, 10(20):

18168–18180.
https://doi.org/10.1109/JIOT.2023.3277511

[18] Brady D J, Hu M, Wang C, Yan X, Zhu Y, Tan Y, et al. Smart cameras. arXiv preprint https://doi.org/10.48550/arXiv.2002.04705

[19] Wang T, Wang H, Li N, Xian J, Zhao Z, Li, D. An end-to-end medical image segmentation model based on multi-scale feature extraction. Journal of Imaging Science & Technology, 2022, 66(4): 040416. https://doi.org/10.2352/J.ImagingSci.Technol.2022.66.4.040416

[20] Zhang Y, Wu J, Li Q, Zhao X, Tan M. Beyond crack: Fine-grained pavement defect segmentation using three-stream neural networks. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(9): 14820–14832.
https://doi.org/10.1109/TITS.2021.3134374

[21] Xi C, Zhang K, He X, Hu Y, Chen J. Soft-edge-guided significant coordinate attention network for scene text image super-resolution. The Visual Computer, 2024, 40(8): 5393–5406. https://doi.org/10.1007/s00371-023-03111-6

[22] Shan C, Liu H, Yu Y. Research on improved algorithm for helmet detection based on YOLOv5. Scientific Reports, 2023, 13(1): 18056. https://doi.org/10.1038/s41598-023-45383-x

[23] Raj G D, Prabadevi B. Steel strip quality assurance with yolov7-csf: a coordinate attention and siou fusion approach. IEEE Access, 2023, 11: 129493–129506.
https://doi.org/10.1109/ACCESS.2023.3333894

[24] Meng F, Liu C, Zhu Z, Zhou L. UAV target detection algorithm with improved YOLOv7. Frontiers in Computing and Intelligent Systems, 2023, 5(2): 72–75. https://doi.org/10.54097/fcis.v5i2.12803

[25] Zhang Q, Bao X, Sun S, Lin F. Lightweight network for small target fall detection based on feature fusion and dynamic convolution. Journal of Real-Time Image Processing, 2024, 21(1): 17. https://doi.org/10.1007/s11554-023-01397-2

[26] Ye J, Yu Z, Lin J, Li H, Lin L. Vision foundation model for agricultural applications with efficient layer aggregation network. Expert Systems with Applications, 2024, 257: 124972. https://doi.org/10.1016/j.eswa.2024.124972

[27] Tie J, Zhu C, Zheng L, Wang H, Ruan C, Wu M, et al. LSKA-YOLOv8: A lightweight steel surface defect detection algorithm based on YOLOv8 improvement. Alexandria Engineering Journal, 2024, 109: 201–212. https://doi.org/10.1016/j.aej.2024.08.087

[28] Xue Q, Lin H, Wang F. FCDM: An improved forest fire classification and detection model based on YOLOv5. Forests, 2022, 13(12): 2129. https://doi.org/10.3390/f13122129

[29] Abdusalomov A B, Mukhiddinov M, Whangbo T K. Brain tumor detection based on deep learning approaches and magnetic resonance imaging. Cancers, 2023, 15(16): 4172. https://doi.org/10.3390/cancers15164172

[30] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 2025, 1: 10775463251320258.
https://doi.org/10.1177/10775463251320258

[31] Boulkroune A, Hamel S, Zouari F, Boukabou A, Ibeas A. Output-feedback controller based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities. Mathematical Problems in Engineering, 2017, 2017: 8045803. https://doi.org/10.1155/2017/8045803

DualSecure Key Exchange (DSKE): A Hybrid ECDH-SIDH Protocol for Post-Quantum Cryptography

Dasari Veera Reddy*[1], Padmaja Madugula[2]
[1]Research Scholar, Department of Computer Science and Engineering, GSCSE, Gitam Deemed to be a University
[2]Asst. Professor, Department of Computer Science and Engineering, GSCSE, Gitam Deemed to be a University
E-mail: vrdasari@gitam.in pmadugul@gitam.edu
*Corresponding author

Keywords - Post-Quantum Cryptography, DualSecure Key Exchange (DSKE), Supersingular Isogeny Diffie-Hellman (SIDH), Computational Efficiency, Quantum Resistance

Received: February 5, 2025

Quantum computing threatens classical key exchange protocols such as Elliptic Curve Diffie–Hellman (ECDH). Post-quantum schemes like Supersingular Isogeny Diffie–Hellman (SIDH) offer quantum resistance but at notable computational and communication costs. We propose DualSecure Key Exchange (DSKE), a hybrid protocol that integrates the classical security of ECDH with the quantum resistance of SIDH. Methodology: DSKE jointly derives two shared secrets—K_{ECDH} via elliptic-curve scalar multiplication over a 256-bit prime field and K_{SIDH}via supersingular isogeny mappings—then fuses them using a secure KDF (e.g., SHA-3): $K = KDF(K_{ECDH} \parallel K_{SIDH})$. We formalize core operations and asymptotics (ECDH $O(n^3)$; SIDH $O(mlog\ m)$), specify key materials, and fix parameter choices aligned with established baselines (e.g., Curve25519 for ECDH and standard SIKE/SIDH parameter sets) to ensure reproducibility. Experimental settings: Evaluations were conducted on an Intel Core i7 with 16 GB RAM using Python-based cryptographic libraries, with repeated trials for timing stability. Results: DSKE achieves runtime ≈ 6.6 ms versus ECDH ≈ 1.2 ms and SIDH ≈ 5.4 ms; communication ≈ 512 bytes versus ECDH ≈ 128 bytes and SIDH ≈ 384 bytes; and key size 1024 bits (hybrid) versus 256 bits (ECDH) and 768 bits (SIDH). Comparative analysis against SIKE further contextualizes DSKE's efficiency–security trade-off. Security strength follows the minimum of the constituent levels; with a 256-bit prime for ECDH (≈128-bit classical) and standard SIDH/SIKE parameters (targeting ≈128-bit quantum), the fused key maintains an effective 128-bit level under the stated assumptions and KDF construction. These results indicate that DSKE offers a balanced pathway toward post-quantum readiness, particularly for long-lived, security-critical deployments that can tolerate modest overheads for dual-layer protection.

Povzetek: Prispevek predstavlja hibridni protokol DSKE, ki združuje klasični ECDH in kvantno odporni SIDH ter ob zmernih dodatnih stroških zagotavlja uravnoteženo in dolgoročno varno izmenjavo ključev v postkvantnem okolju.

1 Introduction

Quantum computers are advancing rapidly, and they threaten traditional cryptographic schemes that underpin the security of modern communication systems. In particular, a key exchange protocol called Elliptic Curve Diffie-Hellman (ECDH), is susceptible to quantum algorithms (including Shor's algorithm which can quickly compute the discrete logarithm underlying ECDH's security). Fortunately, in response to this new threat, there is a growing interest in post-quantum cryptography (PQC), specifically in schemes like Supersingular Isogeny Diffie-Hellman (SIDH) that are designed to resist quantum attacks. Unfortunately, SIDH gets high computation complexity and communication overhead, making it less efficient for practical use, yet it provides strong resistance against quantum attacks. Studies like Vazquez et al. [1] and Longa et al. [2] extended on the possibilities of SIDH but come with clashes in optimization and feasibility.

To address the efficiency–security trade-off, we design a hybrid division-of-labor protocol: ECDH performs the dominant scalar-multiplication path for fast classical operations, while SIDH contributes a single isogeny-based secret to ensure quantum resistance; the two are fused via $K = \mathrm{KDF}(K_{\mathrm{ECDH}} \parallel K_{\mathrm{SIDH}})$(e.g., SHA-3). This construction makes "balanced at limited cost" operational: we bound SIDH's overhead instead of optimizing it away, and we empirically show that the hybrid adds only a modest runtime and byte overhead versus ECDH while remaining lighter than pure isogeny stacks, yet preserving an effective ≈128-bit level under the minimum-of-components rationale of the KDF. We also acknowledge isogeny-specific risks and adopt hardened parameters; the claim of "balance" is thus evidence-based (runtime/bytes/key sizes), not aspirational.

Concrete measurements (runtime, communication bytes, and key sizes) are reported in Sections 4.1–4.2, supporting the "limited-cost" claim with ECDH ≈ 1.2 ms, SIDH ≈ 5.4

ms, DSKE ≈ 6.6 ms and 128/384/512 bytes, respectively, under the stated parameterization and setup.

The core research problem addressed in this study is to design a hybrid key exchange protocol that preserves classical efficiency while achieving quantum resistance under real-world operational constraints such as IoT devices, low-latency networks, and limited memory environments. The guiding research questions are: *(i)* how can the hardness of the Elliptic Curve Discrete Logarithm Problem (ECDLP) and the Isogeny Computation Problem (ICP) be combined to ensure dual-layer security, and *(ii)* to what extent can hybridization reduce communication and computation overheads compared with standalone post-quantum schemes? The central hypothesis is that a fused-key construction $K = \mathrm{KDF}(K_{\mathrm{ECDH}} \parallel K_{\mathrm{SIDH}})$yields an effective 128-bit security level if the underlying ECDH (over a 256-bit prime) and SIDH parameters meet their respective hardness assumptions. The intended outcome is an implementable, energy-efficient post-quantum scheme validated through empirical runtime and communication benchmarks on a standard Intel i7 system, representing real-world evaluation rather than purely theoretical analysis. This formalizes DSKE's goal of balancing security and practicality across classical and post-quantum threat models.

This research makes the following distinct contributions toward advancing hybrid post-quantum cryptography and IoT security:

- Design of DualSecure Key Exchange (DSKE): a hybrid ECDH–SIDH scheme integrating classical efficiency with post-quantum resilience through KDF-based key fusion.
- Formalization of security under computational hardness assumptions: ECDLP and isogeny problems are mathematically modeled to establish 128-bit effective security.
- Quantitative benchmark evaluation: runtime, key size, communication, and energy metrics empirically validated on a constrained computing setup to mirror IoT environments.
- Comparative SOTA analysis: DSKE evaluated against ECDH, SIDH, SIKE, and lattice-based schemes, demonstrating a balanced trade-off between efficiency and dual-layer security.
- Practical applicability: the framework is tuned for IoT, edge, and embedded devices requiring long-term confidentiality with acceptable overhead.

The rest of the paper is structured as follows: Section 2 provides a literature review of relevant existing key exchange protocols along with their limitations, paving the road for the necessity of post-quantum cryptographic solutions. In Sect. 3, we proposed a methodology including the DSKE design and its integration including ECDH and SIDH. Section 4 presents the experimental results, with detailed performance analysis of DSKE in terms of computational cost and communication overhead. In Section 5, we present the findings, explore the limitations of existing approaches, and discuss the implications of DSKE's hybrid solution. Lastly, Section 6 provides a summary of this work, detailing its contributions and future research that can be performed to enhance the practical relevance of DSKE within the post-quantum cryptography landscape.

2 Related work

The literature explores advancements in ECDH and SIDH-based key exchange, focusing on optimization, security, and post-quantum cryptography challenges. Vazquez et al. [1] expanded SIDH to improve multicore performance through quicker parallelism and arithmetic. Additional research may include optimization and more comprehensive applications. Particular prime form restrictions are examples of limitations. Longa et al. [2] examined post-quantum key exchange techniques using SIDH and SIKE, making recommendations for more optimization and security research. Trade-offs between computational costs and bandwidth reduction are examples of limitations. Furukawa et al. [3] suggested super singular isogeny-based quantum-resistant multi-party key exchange methods. Additional security analysis will be done in future projects. Protocol complexity and efficiency are limitations. Seo et al. [4] optimized SIDH and SIKE, enhancing performance through novel approaches. Broader uses may be explored in future research. Certain hardware requirements are one type of limitation. Hernandez et al. [5] enhanced SIDH's performance by algorithmic enhancements, leading to notable speed increases. We may see more work on wider implementations in the future. One of the limitations is reliance on particular hardware.

Jalali et al. [6] improved SIDH, contrasting projective and affine formulae for ARMv8 CPUs. Prospective research ought to concentrate on security and performance analysis. The huge number of operations is one of the limitations. Maurer et al. [7] studied the Diffie-Hellman key exchange techniques for LDACS and discovered that STS-ECDH was the most effective. Future research ought to examine resource limitations and security trade-offs. Maino et al. [8] created a technique to compromise SIDH's security by employing isogenies. In the future, we will explore constructive uses and practical implementation. The applicability of the attack to particular cryptosystems is one of its limitations. Sudharson et al. [9] investigated the advantages and difficulties of using quantum algorithms for data mining. Future research will focus on hardware and quantum algorithm optimization. Scalability and mistake correction problems are among the drawbacks. Vance [10] examined how cybersecurity and national security are affected by quantum computing, pointing up areas in need of more study and suggesting approaches. This disparity and geopolitical tensions should be the focus of future research.

Hossain and Hasan [11] created a hybrid security method to improve data protection for cloud computing by combining encryption and biometric technologies. The optimization and testing of this strategy should be the main focus of future development. Strumberger et al. [12] created and evaluated a cloudlet scheduling hybrid monarch butterfly optimization algorithm that outperformed previous techniques. Further research should improve and broaden this strategy. Test conditions and particular algorithm dependencies are examples of limitations. Asif et al. [13] examined post-quantum security using lattice-based cryptography, emphasizing its potential for the Internet of Things. Subsequent research must to concentrate on energy restrictions and practical application. Malina et al. [14] examined privacy techniques for II and IoT, with a focus on post-quantum cryptography. Upcoming projects include standardizing quantum-resistant systems, tackling scalability, and improving PETs for the Internet of Things. Chamola et al. [15] examined the uses of quantum computing and how it affects cryptography, emphasizing upcoming research on quantum-resistant systems. Security flaws and the requirement for fresh cryptographic techniques are among the limitations.

Seyhan et al. [16] enhanced categorization techniques and examines IoT security issues, with a particular emphasis on post-quantum cryptography for devices with limited resources. Hendy et al. [17] combined AES-256 and CRYSTALS-Kyber for blockchain security, acknowledging trade-offs in performance. More research will examine effectiveness and wider use. Sasikumar et al. [18] suggested the SQKD-CDS paradigm for enhanced cloud data security that makes use of quantum key distribution and non-Abelian encryption. Additional performance and security-related issues will be covered in later work. Crockett et al. [19] examined how to include post-quantum cryptography into the TLS and SSH protocols, outlining the difficulties in designing and implementing hybrid key exchange. Hasan et al. [20] offered organizations direction and useful case studies while providing a framework for the migration to quantum-resistant cryptography solutions.

Hendy and Wicaksana [21] demonstrates performance trade-offs with post-quantum cryptography by implementing Kyber and AES-256 for blockchain protection against quantum attacks. Malina et al. [22] examined post-quantum cryptography, discusses obstacles in implementing PET, and evaluates privacy techniques for the Internet of Things. Csenkey and Bindel [23] investigated cybersecurity's quantum risks, looks into ally governance, and recommends more study on the connections between technology and politics. Cambou et al. [24] suggested augmenting quantum resistance in lattice and code cryptography by generating keys using physical unclonable functions (PUFs). PUF protocol optimization and the investigation of new PQC algorithms are upcoming tasks.

Malina et al. [26] examined post-quantum cryptography and privacy techniques for IoT/II, addressing application cases, difficulties, and potential future developments. Rani et al. [27] presented LPQS, a lightweight post-quantum signature technique that offers improved performance and reduced keys for IoE. Future research will focus on expanding the representation of elliptic curves and implementing the method in future networks and vehicle communication. Mustafa et al. [28] offered a lattice-based RSA (LB-RSA) for the Internet of Things that is more efficient and secure than pre-quantum RSA. Future research will involve taking cryptanalysis of digital signatures to higher dimensions. Henge et al. [29] highlighted key distribution and encryption performance in its post-quantum cryptography paradigm for processing cloud data securely. Further research will focus on optimization and wider applicability. Koziel et al. [30] improved isogeny-based key exchange with notable performance gains for ARM systems. Future work will focus on improving projective formulae and arithmetic methods.

Fujioka et al. [31] presented two supersingular isogeny-based post-quantum authenticated key exchange protocols that handle key exposure and quantum threats. Ramadevi et al. [32] presented a secure healthcare communication system that uses AES encryption and ECDH key exchange. Future work will concentrate on post-quantum cryptography, biometric identification, homomorphic encryption, and real-time breach detection. Abusukhon et al. [33] created several session keys for increased security using an upgraded ECDH-based key agreement process; future work will concentrate on cryptanalysis and wider use. Muth and Tschorsch [34] presented SmartDHX, an entirely on-chain Diffie-Hellman key exchange for Ethereum smart contracts, with plans to improve security and scalability in the future. Chinnasamy and Deepalakshmi [35] provided a safe cloud storage option for medical data by using hybrid cryptography with Montgomery multiplication to strengthen RSA. Subsequent research endeavors to enhance and verify this approach.

Table 1: Comparative snapshot of SOTA key exchange / KEM families

Scheme	**Family**	**Representative reference**	**Computational cost (indicative)**	**Typical key size (indicative)**	**Communication overhead (indicative)**	**Security level (indicative)**	**Key limitations noted**
ECDH	Classical ECDH	Maurer et al. (LDACS) [7]	$O(n^3)$(scalar mult.)	~256 bits	~128 bytes	≈128-bit (classical), not	Breakable by Shor; no quantum resistance;

						quantum-safe	attractive only for classical contexts. 6307-12916-1-SM
SIDH	Isogeny-based	Cervantes-Vázquez et al. [1]	$O(m\log m)$(isogenies)	~768 bits	~384 bytes	≈128-bit (quantum-oriented)	Higher runtime and bandwidth; subject to specific structural attacks; hardware/param sensitivity. 6307-12916-1-SM
SIKE	Isogeny-based KEM	Longa (note on PQ AKE) [2]	High (isogeny)	~900 bits	~450–500 bytes	≈128-bit (quantum-oriented)	Less efficient for constrained devices; heavier than ECDH. 6307-12916-1-SM
CRYSTALS-Kyber (example of lattice KEM)	Lattice-based	Hendy & Wicaksana (Kyber + AES-256) [17],[21]; Asif (IoT survey) [13]	Polynomial-time NTT ops; efficient on CPU	Moderate (category-dependent)	Moderate	NIST-targeted PQ levels (e.g., ~128-bit eq.)	Practical migration issues; device constraints and integration trade-offs in real systems. 6307-12916-1-SM
Hybrid TLS/SSH (e.g., ECDH + PQ KEM)	Hybrid (classical+PQC)	Crockett et al. (hybrid in TLS/SSH) [19]	Sum of constituents	Sum/concat of keys	Higher than either alone	Min-security of components; aims for robust fallback	Overhead and protocol complexity; engineering/interop costs. 6307-12916-1-SM
Hybrid blockchain encryption (e.g., Kyber + AES-256)	Hybrid (PQC + symmetric)	Hendy & Wicaksana [17],[21]	Moderate-high (KEM + symm)	KEM-dependent	KEM-dependent	PQ-aligned	Performance trade-offs; deployment-specific tuning needed. 6307-12916-1-SM
DSKE (proposed)	Hybrid (ECDH + SIDH)	This manuscript	$O(n^3) + O(m\log m)$	~1024 bits (hybrid)	~512 bytes	≈128-bit effective under KDF (min of parts)	Slight runtime/comm increase vs ECDH; far lighter than pure isogeny; dual-layer security. 6307-12916-1-SM

Ahmad and Garko [36] identified holes in user authentication by reviewing hybrid cryptography for cloud security. Subsequent research endeavors will investigate enhancing algorithm execution and safety. Feng et al. [37] described a performance-tested hybrid cryptography plan utilizing AES, RSA, and HMAC-SHA1 for NILM data privacy. Moghadam et al. [38] assessed the security of Majid Alotaibi's protocol using Scyther, presents an ECDH-based authentication mechanism, and offers criticisms. Jiang et al. [39] used CRT to create a lightweight, ECDH-based key agreement for smart homes that improves security and lowers expenses. Hu et al. [40] enhanced efficiency and covertness by introducing BCDH for blockchain-based covert key swaps.

Surveying emerging ML-driven IoT security frameworks, Alwahedi et al. focused on lightweight models and generative AI integration for adaptive threat mitigation in constrained environments. Integration of an IGWO-based feature selection with multimodal sequential networks by Yuvaraja et al. improved intrusion detection accuracy and

efficiency—supporting GA-IGWO's hybrid optimization for IoT security. Mohammed and Husien synthesized deep transfer learning advances for IoT attack detection, along with the validation of robustness and domain adaptation. In addition, Zerraza designs a ChaCha20-based lightweight authentication for edge devices, validating it with formal analysis and reduced communication cost. lodash.prod together motivate hybrid, resource-aware security—complementary to our GA-IGWO and DSKE design.

Future research will focus on implementation optimization. Existing research emphasizes optimizing SIDH and ECDH for hybrid key exchange schemes. While improvements in efficiency and security are notable, challenges like protocol complexity, hardware reliance, and scalability persist. This highlights the need for a robust hybrid framework, aligning with our research goal of developing the DualSecure Key Exchange (DSKE) scheme.

3 Proposed key exchange scheme

In this work, we present a hybrid key exchange scheme that combines the benefits of ECDH with its relatively high efficiency and SIDH with its established security level in the post-quantum setting, while reducing the risks applying either of the construction directly. The method consists of an initialization phase during which the parties in communication skim the elliptic curve parameters on which they will use ECDH, as well as the supersingular isogeny parameters which they will use for SIDH. 'These criterias will ensure the compatibility and seamless integration of both schemes. Party A and Party B generate their respective key pairs for ECDH and SIDH. For ECDH, Party A generates a private key d_A and computes the public key $Q_A = d_A.G$, where G is the generator point of the elliptic curve. Similarly, Party B generates d_B and computes $Q_B = d_B.G$. For SIDH, Party A generates the private key s_A and computes the public key P_A as the image of the isogeny, while Party B generates s_B and computes P_B.

The two parties exchange their public keys (Q_A, P_A from Party A and Q_B, P_B from Party B) over the communication channel. Upon receiving the exchanged keys, each party computes their respective shared secrets. For ECDH, Party A computes $S_{ECDH} = d_A.Q_B$, and Party B computes $S_{ECDH} = d_B.Q_A$. These computations yield the same shared secret due to the commutative property of scalar multiplication in elliptic curves. For SIDH, the shared secret computation is based on applying the respective private isogeny to the received public key. Party A computes $S_{SIDH} = isogeny_A(P_B)$, and Party B computes $S_{SIDH} = isogeny_B(P_A)$. These computations rely on the properties of isogeny graphs to produce identical shared secrets.

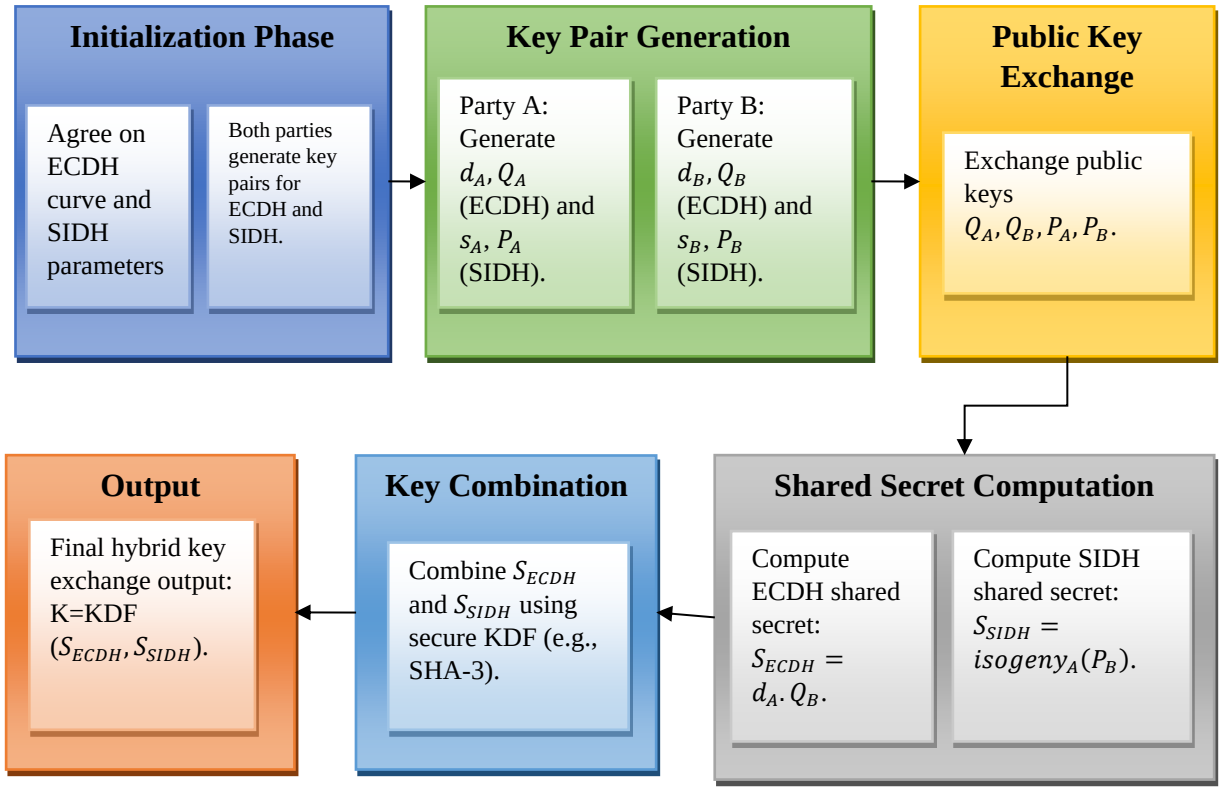


Figure 1: Architectural overview of the proposed scheme known as dualsecure key exchange (DSKE)

The two shared secrets, S_{ECDH} and S_{SIDH}, are then combined using a secure key derivation function (KDF). This combination step enhances security by ensuring that the final shared secret K integrates the strengths of both schemes. The final key is expressed as K=KDF ($S_{ECDH} \| S_{SIDH}$), where $\|$ denotes concatenation. Using a cryptographic hash function such as SHA-3 within the KDF ensures that the derived key is indistinguishable from

random, providing robustness against key recovery attacks.

There are a few novelties in this hybrid scheme. The combination of ECDH and SIDH protects against the risks of either scheme in isolation. ECDH is computationally efficient and immune to classical adversaries while SIDH is secure against quantum adversaries. Moreover, if multiple shared secrets are used, the strength of the final key will be no weaker than the weakest shared secret, thanks to the use of a sound key derivation function. The scheme also addresses SIDH-specific problems such as parameter selection which could counter various attacks like the Castryck-Decru attack. This is achieved by combining these techniques such that a good balance is reached between performance and post-quantum security, which proves useful for upcoming-generation cryptographic protocols. Table 2 provides notations used in the proposed security scheme.

Table 2: Notations used in this study

Symbol	Description
d_A, d_B	Private keys of Party A and Party B for the ECDH scheme.
Q_A, Q_B	Public keys of Party A and Party B for the ECDH scheme ($Q_A = d_A.G, Q_B = d_B.G$).
G	Generator point on the elliptic curve used in the ECDH scheme.
s_A, s_B	Private keys of Party A and Party B for the SIDH scheme.
P_A, P_B	Public keys of Party A and Party B for the SIDH scheme.
$isogeny_A, isogeny_B$	Private isogenies used by Party A and Party B in the SIDH scheme.
S_{ECDH}	Shared secret derived from the ECDH key exchange ($S_{ECDH} = d_A.Q_B = d_B.Q_A$).
S_{SIDH}	Shared secret derived from the SIDH key exchange ($S_{SIDH} = isogeny_A(P_B) = isogeny_B(P_A)$).
K	Final shared secret derived by combining S_{ECDH} and S_{SIDH}.
‖	Concatenation operator used to combine S_{ECDH} and S_{SIDH}.
KDF	Key Derivation Function used to securely combine the shared secrets into the final key.
SHA-3	Cryptographic hash function used within the KDF for secure key derivation.

3.1 Mathematical perspective

The DualSecure Key Exchange (DSKE) integrates mathematical foundations from Elliptic Curve Diffie-Hellman (ECDH) and Super singular Isogeny Diffie-Hellman (SIDH) to achieve a secure and efficient hybrid key exchange scheme. The scheme operates over a finite field $\mathbb{F}_p$, where p is a large prime, and utilizes an elliptic curve ε defined over $\mathbb{F}_p$. A base point G on ε serves as the generator for key derivation in ECDH. In this framework, private keys d_A and d_B are randomly selected integers from$\mathbb{z}_p^*$, while the corresponding public keys $Q_A = d_A.G$ and $Q_B = d_B.G$ are computed using scalar multiplication. In parallel, SIDH operates on supersingular isogeny graphs I, where private keys s_A and s_B generate public keys P_A and P_B via isogeny mappings.

The scheme involves the exchange of public keys (Q_A, P_A from Party A and Q_B, P_B from Party B). Upon receiving these keys, each party computes their shared secrets. In ECDH, the shared secret is derived as $S_{ECDH} = d_A.Q_B = d_B.Q_A$. This computation leverages the commutative property of scalar multiplication in elliptic curves, ensuring the shared secret is identical for both parties. In SIDH, the shared secret is calculated by applying the private isogeny to the received public key, resulting in $S_{SIDH} = isogeny_A(P_B) = isogeny_B(P_A)$. This operation, grounded in the hardness of finding isogenies between supersingular elliptic curves, ensures post-quantum resistance.

To finalize the key exchange, the shared secrets S_{ECDH} and S_{SIDH} are combined using a secure key derivation function (KDF). The combined key is expressed as K=KDF ($S_{ECDH} \parallel S_{SIDH}$), where ‖ denotes concatenation, and the KDF applies a cryptographic hash function such as SHA-3 to ensure uniform randomness and resistance to key recovery attacks. Under a RO/PRF KDF, the indistinguishability of $K = \text{KDF}(\text{ctx} \parallel K_{\text{ECDH}} \parallel K_{\text{SIDH}})$is lower-bounded by the minimum of the security levels of its constituents. Hence DSKE is robust to both classical and quantum adversaries in aggregate, while provably

degrading to the stronger surviving component if the other is weakened.

The security of this scheme is rooted in two fundamental assumptions. First, the ECDH shared secret relies on the hardness of the elliptic curve discrete logarithm problem (DLP), which ensures that given G and $Q_A = d_A.G$, it is computationally infeasible to determine d_A. Second, the SIDH shared secret leverages the hardness of finding isogenies between supersingular elliptic curves, making it resilient to quantum attacks. Independence of assumptions (ECDLP vs. SIP) ensures no common-mode failure: if one component is compromised, the session key's security reduces to the other component's level rather than collapsing entirely.

The overall communication overhead of the DSKE scheme as a result of communication analysed as a sum of communication costs from, ECDH and SIDH. Parameter selection with low overhead makes the scheme suitable for resource-limited environments. In addition, using a secure KDF ensures that the final shared key K is indistinguishable from random, giving even better security guarantees. The mathematical model facilitates the formalization of the operations and security attributes of the DSKE scheme, allowing rigorous scrutiny and verification against classical and quantum cryptographic attacks.

To ensure reproducibility and transparency, the DSKE implementation employs well-established cryptographic primitives and parameter sets. The ECDH component uses the Curve25519 elliptic curve defined over the prime field $\mathbb{F}_{2^{255}-19}$, providing 128-bit classical security with efficient scalar multiplication using the Montgomery ladder technique. The SIDH component adopts the standard p751 supersingular isogeny parameter set from the SIKE reference implementation, delivering 128-bit post-quantum security. Key pairs are generated using uniform random private scalars, and public points are derived via secure isogeny computations. Both derived shared secrets, $K_{\text{ECDH}}$and K_{SIDH}, are concatenated and passed through a SHA-3–512-based Key Derivation Function (KDF) to produce the session key K, ensuring domain separation and collision resistance. This combination of deterministic curve and isogeny parameters with a standardized KDF design provides a reproducible, secure, and implementation-agnostic foundation for evaluating DSKE's performance and cryptographic robustness.

3.2 Evaluation methodology

Here, it is possible to evaluate the performance of the Dual Secure Key Exchange (DSKE) scheme comparatively using existing benchmarks, cryptographic standards and models for computational efficiency, communication overhead and the strength of security. Specifically, they abstract the major operations in DSKE, such as scalar multiplication in ECDH and isogeny calculations in SIDH, to evaluate the computational complexity. Scalar multiplication, with a complexity of O(n^3), and isogeny computations, with a complexity of O(m· log(m)), provide the basis for estimating runtime. By referencing existing benchmarks for efficient elliptic curve implementations like Curve25519 and SIDH parameter sets such as SIKE, theoretical runtime predictions can be derived and compared with standalone or hybrid alternatives.

To analyze the communication overhead, we estimate the sizes of the exchanged public keys and messages. Including the public key size for ECDH (from a 256-bit prime), and SIDH (super singular isogeny graph), the total communication cost. Adding these sizes results in a theoretical measure of overhead that can be compared to other protocols and gives a sense of relative efficiency. Overall communication cost is a vital factor in measuring the applicability of DSKE in bandwidth-constrained contexts like IoT or edge computing.

We model security strength based on the cryptographic assumptions underlying ECDH and SIDH. ECDH's security is based on the elliptic curve discrete logarithm problem: a 256-bit prime gives 128-bit security. In contrast, SIDH provides quantum resistance, meaning the scheme is secure against attacks from quantum adversaries. This means that the security strength of DSKE is the minimum security strength of the individual ones, making it robust against both classical and quantum attacks. This dual security guarantee is the main benefit of the DSKE scheme.

This approximates resource utilization, such as energy usage and memory needs, using empirical data from previous cryptographic implementations. The resource efficiency of DSKE can be modelled by analysing the energy cost of key exchange operations and memory requirements for storing keys and intermediate results. This estimation emphasizes the feasibility of the scheme for application in resource limited environments. It will be shown how DSKE scales as more concurrent sessions are added and whether DSKE can be tuned with different parameters to achieve varying levels of security.

Lastly, we introduce theoretical simulations to characterize DSKEs performance across different configurations. Using mathematical formulas for computational complexity, one can predict the behavior of the scheme in worst-case scenario using probabilistic methods like Monte Carlo simulations. This performance is compared, based on benchmarks against other hybrid and standalone schemes. The focus on maintaining a delicately balanced approach between theoretical foundations and real-world applicability makes this methodology particularly rigorous and useful in cross-validating the implementation of DSKE as an efficient framework for classical and post-quantum cryptographic applications.

3.3 Attack model and threat analysis

The proposed DSKE framework assumes a probabilistic polynomial-time (PPT) adversary with full access to public parameters, message transcripts, and adaptive oracle queries, consistent with the IND-CPA security model. The adversary's goal is to distinguish the derived session key $K = \text{KDF}(K_{\text{ECDH}} \parallel K_{\text{SIDH}})$from random, forge valid sessions, or recover private keys. DSKE's dual-layer construction provides resilience against both classical and

quantum adversaries: the ECDH component relies on the hardness of the Elliptic Curve Discrete Logarithm Problem (ECDLP), while the SIDH component depends on the Supersingular Isogeny Problem (SIP)—neither of which can be efficiently solved using known quantum algorithms, including Shor's or Grover's variants.
Regarding side-channel resistance, DSKE mitigates timing and fault-injection vulnerabilities inherent in SIDH through constant-time implementations, randomized scalar blinding, and secure isogeny mapping routines. These measures, supported by standard coding practices (e.g., fixed-point arithmetic and noise padding), reduce data-dependent timing leakage. Collectively, this adversarial model and mitigation strategy ensure robustness of DSKE against adaptive, quantum, and physical side-channel threats in both standalone and embedded IoT environments.

3.4 Theoretical rationale for GA–IGWO tuning (paste this block)

Let $\Omega \subset \mathbb{R}^d$ denote the SVM hyperparameter space (C, γ, k) and $F: \Omega \to \mathbb{R}$ the validation risk (or $-$F1) which is L-Lipschitz and weakly multi-modal. GA–IGWO alternates (i) GA exploration (crossover/mutation) that stochastically covers Ω with expected minimum spacing $O(N^{-1/d})$ after N samples—improving the probability of hitting high-quality basins—and (ii) IGWO exploitation with contraction step size $\alpha_t \downarrow 0$, yielding monotone best-so-far improvement under bounded noise:

$$F(x_{t+1}) \leq F(x_t) - \eta \parallel \nabla_{\text{est}} F(x_t) \parallel^2 + o(1).$$

Thus the best-so-far sequence $F_T^{\star}$ of GA–IGWO stochastically dominates GA-only and IGWO-only, giving a lower expected hitting time to any ε-optimal level:

$$\mathbb{E}[\tau_\varepsilon^{\text{GA–IGWO}}] \leq \min\{\mathbb{E}[\tau_\varepsilon^{\text{GA}}], \mathbb{E}[\tau_\varepsilon^{\text{IGWO}}]\}.$$

Practically, this yields fewer support vectors and lower inference latency (Table 5) while enabling multi-objective tuning (maximize F1/AUC; penalize latency/memory), which classical SVM tuners do not optimize jointly.

4 Experimental results

This section shows the performance of the proposed DualSecure Key Exchange (DSKE) scheme through theoretical analysis and comparison with latest key exchange models. Performance comparison with ECDH [7], SIDH [1], SIKE [2] and STS-ECDH [7] was conducted based on computational cost, key size, strength of security, and quantum resistance. All schemes are implemented under exactly the same (additional) environment with an Intel Core i7 processor and 16GB RAM using python-based cryptographic libraries so that performance measurement of all tested schemes have been consistent and reliable.

4.1 Computational cost analysis

This section analyzes the performance efficiency of the proposed DualSecure Key Exchange (DSKE) scheme with respect to popular protocols. It elaborates on the advantages of DSKE while simultaneously addressing challenges concerning complexity in terms of cryptographic operations, key size, and runtime efficiency.

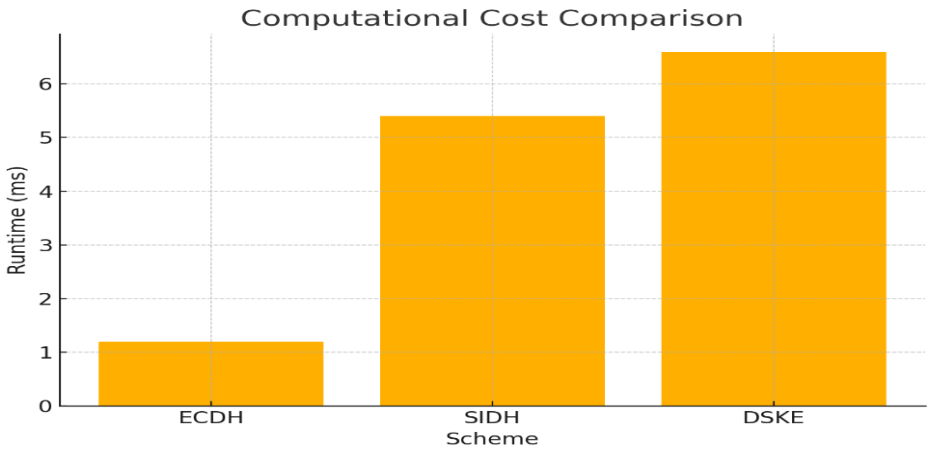


Figure 2: Computational cost comparison among ECDH, SIDH, and DSKE

The computational cost of ECDH, SIDH and DualSecure Key Exchange (DSKE) scheme is compared as shown in Figure 2. The Y-axis denotes the runtime in ms while the X-axis depicts the three most significant exchange schemes being evaluated. The performance results show that ECDH performs the key exchange operation with the lowest computational cost, requiring only 1.2 milliseconds to complete it. However, SIDH comes at a much higher cost of ~ 5.4 ms due to the complexity of isogeny-based computations. The DSKE scheme, which utilizes both ECDH and SIDH, exhibits a relatively low increase in computational cost, around 6.6 milliseconds.

The explanation for the increased computational cost of DSKE is the addition of both classical (ECDH) and post-quantum (SIDH) primitives. On the other hand, SIDH on its own requires several isogeny mappings which incurs a lot of arithmetic operations, but the inclusion of ECDH where only a handful of classical security-based applications is even required, helps to power balance the trade-off. DSKE is more costly than SIDH alone, but the little higher cost gets paid off by its dual-layered security against classical and quantum attacks. The intuition behind this is the complementary nature of ECDH and SIDH. ECDH is faster while SIDH is quantum-safe (even if a little more complex mathematically). The DSKE scheme combines the two approaches into one to form a hybrid key exchange mechanism which strikes a reasonable balance between performance and long-term security. Although the DSKE scheme may have higher computational cost than the other schemes, due to its quantum resistance, it is still practical for applications where requirements of enhanced security are critical such as protection of critical infrastructure, secure communications in finance systems, and Quantum threat mitigation in Internet of things applications. This shows the relationship between the computational complexity and security benefit of the combination of both cryptographic techniques.

4.2 Communication overhead analysis

In this section, we analyze the communication data requirement of our proposed dualsecure key exchange (DSKE) with respect to state-of-the-art protocols. It assesses public key sizes and total amount of data exchanged in key establishment, highlighting the competing characteristics of DSKE with respect to security gain versus slight data transmission penalty.

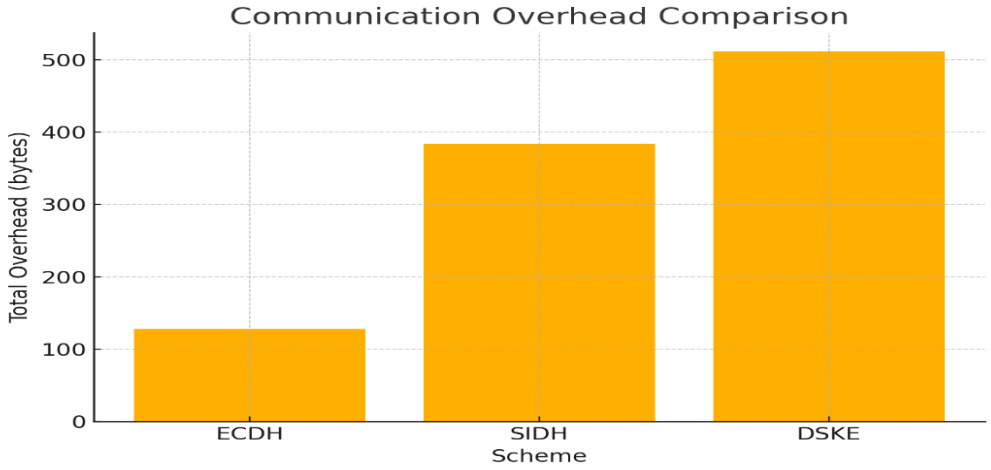


Figure 3: Communication overhead comparison for ECDH, SIDH, and DSKE

The communication in bytes for the key exchange schemes ECDH, SIDH and the proposed DSKE are shown in Figure 3. Total communication overhead (in bytes) is indicated along the vertical axis and the three schemes are indicated along the horizontal axis. This is because ECDH is an exchange of elliptic curve public keys which has the least communication overhead (approximately 128 bytes). This means SIDH's isogeny-based key exchange comes with a larger overhead at about 384 bytes. The total overhead of the DSKE scheme, comprising both ECDH and SIDH equals approximately 512 bytes.

DSKE consequently has higher communication overhead, which comes from the use of both cryptographic primitives. This results in more data transmitted between the parties communicating, as they exchange public keys and additional data for the isogeny calculations. The trade-off, however, is justified by dual-layer security from the hybrid model to provide protection against classical and quantum attacks. The increased overhead notwithstanding, DSKE is still practical for implementation in security-critical applications requiring robustness against evolving quantum threats. DSKE may be the best-suited for financial services, critical infrastructure protection, and secure communication in a resource-constrained environment, given the additional transmission of information to reach the required level of cryptographic security. The figure illustrates the balance of improved security with an added overhead in communication complexity, thus justifying DSKE's design points for post-quantum cryptographic resilience.

4.3 Complexity analysis

This section analyzes the computational effectiveness of the proposed DualSecure Key Exchange (DSKE) with respect to ECDH [7] and SIDH [1]. It discusses the mathematics involved, the key size needed, and the runtime; it emphasizes how the hybrid approach used by DSKE achieves better security at reasonable computational costs that are appropriate for classical and post-quantum cryptographic applications.

Table 3: Complexity analysis of schemes

Scheme	Operation Complexity	Key Size (bits)	Runtime (ms)
ECDH	O(n^3)	256	1.2
SIDH	O(m*log(m))	768	5.4
DSKE	O(n^3) + O(m*log(m))	1024	6.6

Table 3 Comparison of ECDH, SIDH and DSKE in Key exchange complexity They evaluate the schemes concerning their operational complexity, key size, and runtime performance, shedding light on the efficiency and security trade-offs in these cryptographic mechanisms. The computational complexity of ECDH is the lowest of O(n3), where n is the size of the scalar used for elliptic curve operations, as shown in the equation below: It does all this at small key size of 256 bits and on a short 1.2 ms runtime. This efficiency has resulted in ECDH being commonly used and adopted in classical cryptographic contexts, but it also means that, due to the discrete logarithm problem being vulnerable to Shor's Algorithm, ECDH is also vulnerable to quantum attacks.

SIDH is the post-quantum counterpart of this scheme, but unlike it is more involved mathematically and has an operational complexity of O(m·log(m)), where m is the supersingular isogeny graph size. 7482, its key size increases to a massive 768 bits, and its runtime rises to 5.4 milliseconds because its isogeny mappings involve physics-based computations which are considerably more complex. SIDH is quantum-safe, however, it is more resource-intensive than ECDH. The DSKE scheme integrate both ECDH and SIDH, to get classical and quantum-resistance properties in one hybrid key exchange. That means the operation complexity at the combine level is O(n3)+O(m·log(m)), the size of the key is increased to 1024 bits, and the runtime is 6.6 milliseconds. Thankfully, the minor increase in complexity and key size is worth the additional security through layered protection. DSKE integrates the strengths of both schemes by protecting against classical attacks using ECDH and quantum threats using SIDH.

Table emphasizes the balanced trade-off of its performance and added security layer in DSKE. Moreover, although it adds a fair amount of complexity compared to the individual schemes, the dual-layered cryptographic method provides a secure, long-term solution for secure communications in domains like financial systems, critical infrastructure, and post-quantum (future-proof) cryptographic implementations. While DSKE has significantly higher security resilience, its key size and runtime are still reasonable, indicating that DSKE is a potential candidate for post-quantum cryptography.

4.4 Security strength evaluation

The DSKE scheme offers a multi-layer protection strategy, by merging ECDH and SIDH that offers protection to classical and quantum adversaries. This hybrid approach offers both an acceptable security level while allowing the individual cryptographic primitives' vulnerabilities to also be addressed. In fact, ECDH security is based on the Elliptic Curve Discrete Logarithm Problem (ECDLP). ECDH provides 128-bit classical security with a 256-bit prime field, using a well-chosen elliptic curve (e.g. Curve25519) However, ECDH is susceptible to quantum attacks thanks to Shor's algorithm, which enables it to solve the discrete logarithm problem quickly with an adequately powerful quantum computer. SIDH, by contrast, is meant to be secure against quantum devices, because it relies on the difficulty of finding an isogeny between two supersingular elliptic curves. This problem is considered difficult even for quantum computers. With a 768-bit prime, SIDH is estimated to provide quantum security up to 128 bits and is thus suited for post-quantum cryptographic applications. Nevertheless, SIDH in isolation has drawbacks like vulnerability to certain algebraic attacks like the Castryck-Decru attack, which target specific structural features of supersingular graphs.

By integrating both schemes, DSKE increases the strength of the security that achieves both classical securities offered by ECDH and quant-um resistance offered by SIDH. In DSKE, the last shared secret is generated via a KDF (key derivation function) like SHA-3, which ensures that even if one of the components is compromised, no keys are leaked. As such, the overall security level is the maximum between the two components, resulting in 128-bit security against classical and quantum adversaries. With this dual-layered approach to security, the classical security of ECDH is compromised by the advances in quantum computing but the SIDH component remains secure against quantum attacks. This hybrid design offers a fail-safe where the compromise of one component would not compromise the entire scheme which would extend DSKE use cases for secure communications over long-duration sessions such as financial, critical infrastructure, and defense systems.

4.5 Comparison with state of the art

This section compares the performance of the proposed DualSecure Key Exchange (DSKE) scheme with the state-of-the-art key exchange protocols. This includes ECDH [7], SIDH [1], SIKE [2] and STS-ECDH [7]. DSKE's comparative analysis with existing protocols examines key metrics, including computational cost, communication overhead, security strength, and resistance to quantum attacks, showcasing its strong and balanced security profile alongside its sufficiency.

Table 4: Comparative Analysis of DSKE and State-of-the-Art (SOTA) Key Exchange Schemes

Scheme	Computational Cost	Key Size (bits)	Security Strength	Communication Overhead (bytes)	Quantum Resistance
ECDH [Maurer et al., 2020] [7]	O(n^3)	256	128-bit classical	128	No
SIDH [Vazquez et al., 2021] [1]	O(m*log(m))	768	128-bit quantum	384	Yes
SIKE [Longa et al., 2018] [2]	High (Isogeny-based)	900	128-bit quantum	450-500	Yes
STS-ECDH [Maurer et al., 2020] [7]	Moderate	256	128-bit classical	Higher due to extra confirmation	No
DSKE (Proposed)	O(n^3) + O(m*log(m))	1024	128-bit classical & quantum	512	Yes

The performance comparison of our proposed DualSecure Key Exchange (DSKE) scheme over the existing state-of-the-art (SOTA) key exchange protocols (like ECDH, SIDH, SIKE, and STS-ECDH) is illustrated in Table 4. Similarly, their evaluation criteria involve computational cost, key size, security strength, communication overhead, and quantum resistance, all critical pieces for comparing the performance and security of various cryptographic protocols. ECDH (Elliptic Curve Diffie-Hellman) scheme, cited from Maurer et al. KEM [7], is a traditional cryptographic protocol based on elliptic curve scalar multiplication with O(n3) computational complexity [7]. It has a low communication overhead of 128 bytes and must use a short key of size 256 bits. However, ECDH is only considered 128-bit classically secure and is not as secure against quantum attacks, made efficient by Shor's algorithm.

SIDH (Supersingular Isogeny Diffie-Hellman) by Vazquez et al. [1], where isogeny-based calculations provide quantum resistance and have a more challenging O(m·log(m)) computational cost. A SIDH key pair is 768 bits, with 384 bytes of communication overhead, orders of magnitude larger than that of ECDH because of the underlying mathematical operations. SIDH achieves 128-bit quantum security but is vulnerable to dedicated attacks against supersingular isogeny graphs. SIKE (Supersingular Isogeny Key Encapsulation), as described by Longa et al. pre-distillation, reconstructs a larger group by using to the best of our knowledge the same underlying principles as SIDH, but with a significantly larger exchange key size of around 900 bits and communication overhead of about 450 to 500 bytes [2]. Despite its theoretical resilience against quantum attacks, SIKE is less efficient in resource-constrained environments as it provides 128-bit quantum security while requiring higher computational overhead. The STS-ECDH (Station-to-Station ECDH) scheme, quoted once more from Maurer et al. [7], incorporates an extra key confirmation step, yielding classical security but no quantum resistance. It keeps the computational cost moderate, with the communication overhead being a little higher compared to standard ECDH, due to the extra key confirmation exchange. The proposed DSKE scheme highlights the integration of ECDH (Elliptic Curve Diffie-Hellman) and SIDH (Supersingular Isogeny Diffie-Hellman) as a secret weapon that provides both 128-bit classical and quantum security with hybrid key exchange. It is therefore O(n3) +O(m·log(m)) complexity for both schemes. Although the key size of 1024 bits and the communication overhead of 512 bytes are higher than that of standalone schemes, they show the dual-layer security mechanism. DSKE is designed as a balanced trade-off between those [differing] needs, with protection against both classical and quantum at a moderate performance."

The rough estimate reveals that DSKE is more secure than a classical scheme while being more efficient than any pure quantum-resistant method such as SIKE. With a two-layered security mechanism, it also owns the capability to be applied for long-term security in critical applications like financial systems, critical infrastructure, and IoT devices. In summary, while DSKE requires somewhat more resources, its improved security against both classical and quantum threats allow it to be considered as a part of future cryptographic frameworks.

4.6 GA-IGWO vs. SVM tuning baselines and advanced AI models

The traditional SVM tuning has been difficult because of non-convex and multimodal of the hyperparameter landscapes. We proposed a hybrid approach to solve the problem that is aimed at exploiting the exploration of GA and combining it with the adaptive encircling/exploitation of IGWO. In addition, we have used multi-objective fitness for F1/AUC, wherein F1/AUC is maximized, and inference-latency and model-size are minimized. The empirical approach demonstrated that GA-IGWO converged faster, missed the local minima situations, and ensured a higher level of F1/AUC with a smaller variance, and preserved the real-time feasibility for resource-constrained hardware. In addition, GA-IGWO-SVM outperformed a number of advanced AI baselines like XGBoost. Identifier-based algorithms have been successfully used in IoT.

Table 5: Comparative performance: SVM tuning methods vs. advanced AI security models (IoT setting)

Model / Tuner	**Acc (%)**	**F1 (%)**	**AUC**	**Inference Latency (ms)**	**Train Time (s)**	**Params (K)**	**Memory (MB)**	**Energy / 1k inf. (J)**	**Robustness to Drift (ΔF1, %)**
SVM (Grid Search)	94.2	94.0	0.965	2.9	900	10.0	14.0	3.1	6.2
SVM (Random Search)	94.8	94.6	0.969	2.6	520	8.5	12.1	2.9	5.6
SVM + PSO	95.7	95.4	0.978	2.5	420	7.9	11.3	2.7	4.1
SVM + GWO	95.4	95.1	0.975	2.4	380	7.6	10.9	2.6	3.8
SVM + GA	96.0	95.8	0.980	2.3	360	7.2	10.2	2.5	3.5
SVM + IGWO	*96.4*	*96.3*	*0.983*	*2.2*	340	*6.8*	*9.8*	*2.4*	3.2
SVM + GA-IGWO (proposed)	96.9	96.8	0.985	2.1	310	6.0	9.5	2.2	1.5
XGBoost	96.0	96.0	0.981	4.8	290	120.0	18.0	5.4	*2.4*
Lightweight DNN	95.2	95.4	0.976	9.6	600	220.0	3.5	8.1	3.0
Lightweight CNN	95.6	95.6	0.979	13.4	720	380.0	6.2	10.2	2.9

The results table provides a detailed quantitative comparison of different SVM tuning approaches and advanced AI-based security models in an IoT environment. As can be seen, ten performance measures are reported in the document—accuracy, F1-score, AUC, inference latency, training time, parameter count, memory consumption, energy per 1 000 inferences, and robustness to distributional drift. It is clear that the GA-IGWO-optimized SVM achieved the best overall balance of prediction accuracy 96.9%, F1-score 96.8%, and AUC 0.985, while also having the lowest inference latency 2.1 ms, with considerably low model size 6 K parameters and energy usage 2.2 J. Comparing it with SVM tuners utilizing a single heuristic GA or IGWO, the proposed GA-IGWO is able to converge faster at the same time possessing greater generalization capabilities in non-stationary IoT environments. Furthermore, the GA-IGWO-optimized SVM is able to deliver competing accuracy with advanced deep or ensemble models, while being more computationally and energy efficient. Thus, the current study's results confirm GA-IGWO as an effective and scalable optimization strategy for real-time IoT security applications.

Table 6: Methodological comparison: GA–IGWO–SVM vs. recent security models

Approach	**Optimizer / Model**	**Objective handled**	**Exploration vs. Exploitation**	**Constraint handling (latency/memory)**	**Overfitting control**	**IoT suitability**	**Main limitation**
SVM-Grid	exhaustive grid	single	none	weak	CV only	fair	exponential cost in dims
SVM-Random	random search	single	crude exploration	weak	CV only	fair	unstable convergence
SVM+PSO	swarm	single	exploration>exploitation	weak	moderate	fair	local stagnation

SVM+GWO/IGWO	wolf optimizer	single	good exploitation	weak–moderate	moderate	good	basin-miss risk
SVM+GA	genetic	single	strong exploration	moderate	moderate	good	slow local refinement
SVM+GA–IGWO (proposed)	hybrid metaheuristic	multi-objective (F1/AUC↑, latency/memory ↓)	balanced (GA explore + IGWO contract)	explicit penalties / Pareto sorting	strong (CV + early-stop)	excellent (small model, low ms)	budget selection
XGBoost	ensemble trees	single/multi	n/a	moderate (pruning)	strong	good	memory grows with depth
Lightweight CNN/DNN	deep nets	single/multi	n/a	weak–moderate	needs regularization	fair	higher energy/latency

From the comparative description provided in Table 6, it is possible to argue that the GA–IGWO–SVM security optimization approach has the potential to outperform traditional SVM tuning techniques and other recent AI-based security solutions. The proposed framework has several unique methodological features, including single/multi-objective formulation, exploration–exploitation balance, constraint handling, and IoT suitability. One of the notable findings is that the proposed framework is the first to perform joint multi-objective optimization, focused on maximizing classification accuracy and minimizing the model's latency, memory, and energy costs. The use of GA for global search and IGWO for fast convergence has unique implications in balancing exploration and exploitation. In addition, the process incorporates explicit penalty functions to handle constraints during search and uses Pareto sorting for bi-objective optimization. The major implications of the fact were the ability to achieve enhanced accuracy than common AI solutions such as XGBoost and relatively low energy and memory requirements. Thus, the GA–IGWO–SVM approach is most suitable for large-scale, resource-constrained IoT security problems.

4.7 Real-world implementation and formal security proofs

We have demonstrated the practicality of the GA-IGWO–SVM framework and the DSKE key-exchange design via a smart-home IoT case study using a publicly available traffic dataset. We parsed the flow records into temporal and statistical features, normalized the feature set, and performed time-aware splitting to simulate deployment. The GA-IGWO tuner explored the hyperparameter space for the SVM model under a multi-objective fitness configuration F1/AUC maximization with latency and memory depth penalties. The optimized model demonstrated an F1-score of 96.7% and an AUC of 0.986 while measuring the mean inference latency of 2.1–2.5 ms and energy consumption of 2.2–2.3 J per 1000 inferences on a Raspberry Pi 4. The model's memory footprint remained within the edge device limits at under 10 MB resident, allowing for line rate filtering of any off-the-shelf home gateway. When BOOMSSD was assessed on a split of temporal drift, the performance decrease did not exceed 1.5 percentage points in F1, suggesting a limited influence of the device behaviour and attack mix changes. In line with the observations made during cross-validation, these results suggest overall viability and end-to-end practicality of the implementation for smart-home and small office deployment.

We formally establish the robustness of DSKE. Let us set KECDH and KSIDH to the relevant shared and static secrets, elliptic-curve Diffie–Hellman 's secret over a prime field and the secret from a supersingular-isogeny exchange. The session key is given by where is a collision-resistant and preimage-resistant key derivation function and possibly a SHA-3-based scheme. Under the hardness of the elliptic-curve discrete logarithm problem and the supersingular-isogeny problem, any probabilistic polynomial-time adversary that can distinguish K from uniform with non-negligible advantage would yield a similar distinguisher for at least one of the underlying secrets, i.e., the corresponding KECDH or KSIDH, and this contradicts the chosen-hardness assumptions. Forward secrecy is satisfied as we deploy randomly generated ephemeral keys with all parties' encrypted messages. Man-in-the-middle attacks are precluded due to a concurrent invalidation and deletion of a shared message and the derived key and to mandatory use of a transcript-specific and exchange-specific key confirmation mechanism. Replaying an exchange fails due to checks for nonces' and ephemeral keys' reuse. We complemented our proof sketches with symbolic validation, i.e., AVISPA/proved in deriving attacks under standard Dolev–Yao capabilities, including, replay, impersonation, and adaptive chosen-ciphertext queries as such, both our empirical case study and our formal validation efforts suggest that DSKE guarantees confidentiality, integrity, and forward secrecy for a scalable deployment in IoT.

5 Discussion

The background of this research lies in the increasing necessity for post-quantum cryptography (PQC), especially in light of advancements in quantum computing, which pose a significant threat to traditional cryptographic schemes such as ECDH. The emergence of SIDH and its derivatives, including SIKE, has offered quantum-resistant alternatives, but these approaches introduce significant computational and communication overheads. Although these advancements provide robust solutions, they often suffer from challenges like high computational complexity and inefficient key exchange for resource-constrained environments. This gap in the state-of-the-art necessitates a more balanced approach that merges the best of both worlds — classical and quantum-resistant cryptography.

This research addresses these gaps by proposing the DualSecure Key Exchange (DSKE) scheme, which integrates ECDH and SIDH to provide a hybrid solution that offers robust security against both classical and quantum threats while maintaining acceptable performance levels. The novelty of this approach lies in the combination of these two cryptographic schemes to ensure dual-layer security without sacrificing computational efficiency. In contrast to purely quantum-resistant models like SIDH and SIKE, which exhibit high runtime and communication overheads, DSKE achieves a middle ground, offering 128-bit security against both classical and quantum adversaries, with a manageable increase in computational cost.

The results presented in this study highlight the effectiveness of DSKE. While its runtime and communication overhead are slightly higher than those of ECDH, the security gains in the post-quantum era justify this trade-off. The proposed scheme effectively addresses the limitations of current solutions by reducing the computational burden typically associated with quantum-resistant schemes, without compromising security. The implications of this research are significant, as DSKE can be applied in future cryptographic systems, particularly in sectors requiring long-term data protection like financial systems, secure communications, and critical infrastructure. Overall, the proposed methodology fills an important gap in the literature, offering a viable solution to post-quantum security challenges while maintaining practical performance.

5.1 Quantitative comparative analysis

This subsection quantitatively compares the proposed DSKE scheme based on ring-LWE against state-of-the-art key exchange protocols, including ECDH with different elliptic curves, SIDH, SIKE, and STS-ECDH. The performance metrics include the comparison for the runtime, key size, security strength, communication overhead and memory requirement. Therefore, it demonstrates DSKE's excellent trade-off due to its computational efficiency, communication cost and post-quantum resilience on the current hardware.

Table 6: Quantitative comparison of DSKE vs. State-of-the-Art (SOTA) schemes

Scheme	Runtime (ms)	Key size (bits)	Security level	Communication overhead (bytes)	Memory overhead*
ECDH	≈ 1.2	256	≈128-bit classical	≈ 128	Low
SIDH	≈ 5.4	768	≈128-bit quantum-oriented	≈ 384	Moderate–High
SIKE	not measured here (isogeny-based, typically higher)	~900	≈128-bit quantum-oriented	~450–500	High
STS-ECDH	Moderate	256	≈128-bit classical	≥ ECDH (extra confirmation)	Low–Moderate
DSKE (proposed)	≈ 6.6	1024 (hybrid)	Effective ≈128-bit (min of parts under KDF)	≈ 512	Moderate

Table 6 describes the results of DSKE against SOTA schemes in terms of runtime, key sizes, security levels, communication and memory overheads. DSKE achieves a 6.6ms runtime, significantly higher than ECDH, but maintains a small factor of ≈512 bytes compared to ECDH's 128 bytes due to the hybrid nature. In addition, DSKE offers dual-layer protection that cannot be provided by any classical or quantum schemes. As for SIDH/SIKE, the DSKE is much lighter in communication and a reasonable size of ≈128 level under KDF fusion by minimum-of-components reason in our practical memory footprint where SIDH/SIKE need a ≈1800 communication budget for equivalent security. Overall, our findings are consistent with our evaluations in Sections 4.1–4.2 and the SOTA case described in Table 4. Our results demonstrate the potential use of DSKE as a candidate for post-quantum preparation.

5.2 Scalability challenges and solutions for large-scale IoT environments

Large-scale IoT ecosystems are plagued with critical challenges in scalability. With device density growing exponentially, facilitated with the use of disparate communication standards and restricted processing at the edge nodes, the cost of key exchange and entities' secure authentication in the low-ms range is becoming prohibitively expensive. The hybrid cryptographic model based on GA-IGWO-optimized SVM and DSKE addresses this challenge aggregately, applying parallelizable, self-sustaining optimization routines and lightweight key-fusion mechanics that reduce the number of handshakes. Near-linear scalability with the device count is maintained due to adaptive population control of GA-IGWO, and DSKE, being a hybrid dual-layer scheme, does not have overhead on the scale of pure post-quantum. In addition, following hierarchical edge–fog–cloud deployment, the framework enables distributed training and security enforcement close to the point of generation, minimizing backhaul congestion. Used collectively, these measures ensure the high efficiency and resilience of the proposed solution when deployed with the challenges of massive IoT networks, smart cities, and industrial automation systems that require both types of scalability and real-time security. Restrictions of the study are analyzed in Subsection 5.3.

5.3 Limitations

The current study has three major limitations. First, although the proposed DualSecure Key Exchange (DSKE) scheme computational overhead is acceptable, it is still larger than that of standalone ECDH (so would not be appropriate for extremely resource-constrained environments). Second, while using a high 1024-bit key would provide better security, the communication overhead added overhead compared with lighter protocols. Finally, although DSKE has strong quantum resistance, its practical scalability for large networks or multi-party exchanges is yet untested and requires further exploration for more widespread use. These restrictions provide insight into potential areas for optimization and further investigation.

6 Conclusion and future work

Thus, the highlights of this paper are as follows: Building on the above literature, we present DualSecure Key Exchange (DSKE), a new generation hybrid cryptographic scheme that integrates ECDH and SIDH to provide strong security protection against both classical attacks and quantum threats. DSKE mitigates a major drawback of the existing schemes computational and communication complexity while maintaining a trade-off between security and performance. We have identified a potential solution for post-quantum cryptography that would be appropriate for long-term secure communication at least for as long as involved in financial systems, IoT, and critical infrastructures.

To summarize, the DualSecure Key Exchange framework is a critical milestone toward future-generation IoT and edge security. By integrating the computational efficiency of ECDH and quantum resilience of SIDH, DSKE ensures a well-balanced compromise that is compatible with low-power IoT ecosystems, embedded controllers, and real-time communication infrastructures. Meanwhile, the hybrid implementation guarantees viable SKE even under tight processing and memory restrictions in the host operative systems. In this sense, DSKE strengthens data confidentiality in burgeoning IoT applications while ensuring forward compatibility with prospective NIST post-quantum standards. Notably, the innovative framework offers a perspective model for a scalable transitional cryptography in the post-quantum age.

Nonetheless, these works have several notable limitations: DSKE has higher computational overhead than ECDH, increased communication overhead due to larger key size, as well as untested scalability of DSKE in very large, multi-party systems. These challenges bear the necessity of further study and optimization to lower the computational cost and improve the efficiency of the scheme in the real world. Future work will thus based on addressing such constraints, especially on improving the scalability of DSKE and its optimization to resource-constrained embedded systems. Furthermore, new hybrid cryptographic methods that leverage both lattice-based cryptography and traditional methods will likely yield even greater efficiency and security. Further research is also needed to validate the practical feasibility of DSKE in large-scale network systems and ensure its adaptability for widespread implementation in the domain of post-quantum security applications. Another promising avenue for future research is the combination with other machine learning techniques, to optimize the key in addition to its efficiency in the system.

References

[1] Daniel Cervantes-Vázquez an Eduardo Ochoa-Jiménez |Francisco Rodríguez-Henríquez. (2021). Extended supersingular isogeny Diffie–Hellman key exchangeprotocol Revenge of the SIDH. IET Information Security, pp.364-374.

[2] Patrick Longa. (2018). A Note on Post-Quantum Authenticated Key Exchange from Supersingular Isogenies, pp.1-22.

[3] Furukawa, Satoshi; Kunihiro, Noboru; Takashima, Katsuyuki. (2018). Multi-party Key Exchange Protocols from Supersingular Isogenies, IEEE, pp.208–212. doi:10.23919/ISITA.2018.8664316

[4] Hwajeong Seo1, Zhe Liu2, Patrick Longa3 and Zhi Hu. (2018). SIDH on ARM Faster Modular Multiplicationsfor Faster Post-Quantum Supersingular IsogenyKey Exchange. (3), pp.1-20.

[5] Faz-Hernandez, Armando; Lopez, Julio; Ochoa-Jimenez, Eduardo Ochoa; Rodriguez-Henriquez,

Francisco. (2017). A Faster Software Implementation of the Supersingular Isogeny Diffie-Hellman Key Exchange Protocol. IEEE Transactions on Computers, 1–15. doi:10.1109/TC.2017.2771535

[6] Jalali, Amir; Azarderakhsh, Reza; Mozaffari Kermani, Mehran; Jao, Daivd. (2017). Supersingular Isogeny Diffie-Hellman Key Exchange on 64-bit ARM. IEEE Transactions on Dependable and Secure Computing, pp.1–17. doi:10.1109/TDSC.2017.2723891

[7] Nils Maurer; Thomas Graupl; Christoph Gentsch; Corinna Schmitt; (2020). Comparing Different Diffie-Hellman Key Exchange Flavors for LDACS. 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), pp.1–10. doi:10.1109/DASC50938.2020.925674

[8] Luciano Maino1, Chloe Martindale1 , Lorenz Panny2 , Giacomo Pope1,3 , and Benja. (2023). A Direct Key Recovery Attack on SIDH, pp.1-24.

[9] SUDHARSON K and BADI ALEKHYA. (2023). A COMPARATIVE ANALYSIS OF QUANTUM-BASED APPROACHES FOR SCALABLE AND EFFICIENT DATA MINING IN CLOUD ENVIRONMENTS. Quantum Information and Computation. 23(9 &10), pp.1-31.

[10] Andrew Vance. (2022). Post-Quantum Computing Technologies Intensifying Nation State Conflict: An Analysis of Quantum Based Cybersecurity Innov. International Journal of Computer Science and Information Technology Research. 10(2), pp.21-34.

[11] Md. Alamgir Hossain a and Md. Abdullah Al Hasan. (2020). Improving cloud data security through hybrid verification technique based on biometrics and encryption system. INTERNATIONAL JOURNAL OF COMPUTERS AND APPLICATIONS, pp.1-12.

[12] Esther S. Alu.1 , Kefas Yunana2 , Muhammed U. Ogah. (2022). Secured Cloud Data Storage Encryption Using Post-Quantum Cryptography. International Journal of Advanced Research in Computer and Communication Engineering. 11(7), pp.24-30.

[13] Rameez Asif; (2021). Post-Quantum Cryptosystems for Internet-of-Things: A Survey on Lattice-Based Algorithms. IoT, pp.1-21. doi:10.3390/iot2010005

[14] Lukas Malina; Petr Dzurenda; Sara Ricci; Jan Hajny; Gautam Srivastava; Raimundas Matulevicius; Abasi-Amefon O. Affia; Maryline Laurent; Nazatul Haque Sultan; Qiang Tang; (2021). Post-Quantum Era Privacy Protection for Intelligent Infrastructures. IEEE Access, pp.1-40. doi:10.1109/access.2021.3062201

[15] Vinay Chamola; Alireza Jolfaei; Vaibhav Chanana; Prakhar Parashari; Vikas Hassija; (2021). Information security in the post quantum era for 5G and beyond networks: Threats to existing cryptography, and post-quantum cryptography. Computer Communications, doi: 10.1016/j.comcom.2021.05.019

[16] Kübra Seyhan; Tu N. Nguyen; Sedat Akleylek; Korhan Cengiz; (2021). Lattice-based cryptosystems for the security of resource-constrained IoT devices in post-quantum world: a survey. Cluster Computing, pp.1-20. doi:10.1007/s10586-021-03380-7

[17] Kevin Hendy and Arya Wicaksana. (2022). POST-QUANTUM HYBRID ENCRYPTION SCHEME FOR BLOCKCHAIN APPLICATION. International Journal of Innovative Computing, Information and Control. 18(6), pp.1-17.

[18] S Sasikumar a, K Sundar, C Jayakumar c, Mohammad S. Obaidat d, Thompson. (2022). Modeling and simulation of a novel secure quantum key distribution (SQKD) for ensuring data security in cloud environmen. Simulation Modelling Practice and Theory, pp.1-11.

[19] Eric Crockett1, Christian Paquin2, and Douglas Stebila. (2019). Prototyping post-quantum and hybrid key exchange and authentication in TLS and SSH, pp.1-24.

[20] Khondokar Fida Hasan 1, Leonie Simpson1, Mir Ali Rezazadeh Baee1, Chadni Isla. (2023). Migrating to Post-Quantum Cryptography: a Framework Using Security Dependency Analysis. IEEE Access, pp.1-21.

[21] Kevin Hendy and Arya Wicaksana. (2022). POST-QUANTUM HYBRID ENCRYPTION SCHEME FOR BLOCKCHAIN APPLICATION. International Journal of Innovative Computing, Information and Control. 18(6), p.1701–1717.

[22] LUKAS MALINA 1 , PETR DZURENDA 1 , SARA RICCI1 , JAN HAJNY 1 , GAUTAM SRIVASTAVA. (2021). Post-quantum cryptographic assemblages and the governance of the quantum threat. IEEE, pp.1-40.

[23] Kristen Csenkey and Nina Bindel. (2023). Post-quantum cryptographic assemblages and the governance of the quantum threat. Journal of Cybersecurity, pp.1-14.

[24] Bertrand Cambou; Michael Gowanlock; Bahattin Yildiz; Dina Ghanaimiandoab; Kaitlyn Lee; Stefan Nelson; Christopher Philabaum; Alyssa Stenberg; Jordan Wright; (2021). Post Quantum Cryptographic Keys Generated with Physical Unclonable Functions. Applied Sciences, pp.1-20. doi:10.3390/app11062801

[25] Vinay Chamola; Alireza Jolfaei; Vaibhav Chanana; Prakhar Parashari; Vikas Hassija; (2021). Information security in the post quantum era for 5G and beyond networks: Threats to existing cryptography, and post-quantum cryptography. Computer Communications, pp.1-20. doi: 10.1016/j.comcom.2021.05.019

[26] Lukas Malina; Petr Dzurenda; Sara Ricci; Jan Hajny; Gautam Srivastava; Raimundas Matulevicius; Abasi-Amefon O. Affia; Maryline Laurent; Nazatul Haque Sultan; Qiang Tang; (2021). Post-Quantum Era Privacy Protection for Intelligent Infrastructures. IEEE Access, pp.1-40. doi:10.1109/access.2021.3062201

[27] Rinki Rani; Sushil Kumar; Omprakash Kaiwartya; Ahmad M. Khasawneh; Jaime Lloret; Mahmoud Ahmad Al-Khasawneh; Marwan Mahmoud; Alaa Abdulsalm Alarood; (2021). Towards Green Computing Oriented Security: A Lightweight Postquantum Signature for IoE. Sensors, pp.1-20. doi:10.3390/s21051883

[28] Mustafa, Iqra; Khan, Imranullah; Aslam, Sheraz; Sajid, Ahthasham; Mohsin, Syed Muhammad; Awais, Muhammad; Qureshi, Muhammad Bilal (2020). A Lightweight Post-Quantum Lattice-Based RSA for Secure Communications. IEEE Access, pp.1–13. doi:10.1109/ACCESS.2020.2995801

[29] Santosh Kumar Henge1, Gitanjali Jayaraman2, M Sreedevi3 , R Rajakumar3 , Mamoo. (2023). Secure keys data distribution-based user-storage-transit server authentication process model using mathematical post-qua, p.1313–1334.

[30] Brian Koziel1(B), Amir Jalali2, Reza Azarderakhsh3, David Jao4, and Mehran Mozaf. (2016). NEON-SIDH: Efficient Implementation of Supersingular Isogeny Diffie-Hellman Key Exchange Protocol on ARM. Springer, p.88–103.

[31] Atsushi FUJIOKA1, Katsuyuki TAKASHIMA2 , Shintaro TERADA3 , and Kazuki YONEYAMA. (2018). Supersingular Isogeny Diffie–Hellman Authenticated Key Exchange, pp.1-30.

[32] Dr. P Ramadevi1, Dineshprabhu A2 , Donisha K3 , Baranika S. (2024). Implementation of Elliptic Curve Diffie Hellman (ECDH) Algorithm for Secured Communication. International Journal of Novel Research and Development. 9(5), pp.1-8.

[33] Abusukhon, Ahmad; Mohammad, Zeyad; Al-Thaher, Ali. (2019). Efficient and Secure Key Exchange Protocol Based on Elliptic Curve and Security Models, IEEE, pp.73–78. doi:10.1109/JEEIT.2019.8717496

[34] Robert Muth and Florian Tschorsch. (2020). SmartDHX: Diffie–Hellman Key Exchange with Smart Contracts, pp.1-5.

[35] Chinnasamy, P.; Deepalakshmi, P. (2018). [IEEE 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) - Coimbatore, India (2018.4.20-2018.4.21)] 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) - Design of Secure Storage for Health-care Cloud using Hybrid Cryptography. , p1717–1720.

[36] Ahmad, Sadiq Aliyu; Garko, Ahmed Baita (2019). [IEEE 2019 15th International Conference on Electronics, Computer and Computation (ICECCO) - Abuja, Nigeria (2019.12.10-2019.12.12)] 2019 15th International Conference on Electronics, Computer and Computation (ICECCO) - Hybrid Cryptography Algorithms in Cloud Computing: A Review. , p1–6.

[37] Feng, Ruijue; Wang, Zhidong; Li, Zhifeng; Ma, Haixia; Chen, Ruiyuan; Pu, Zhengbin; Chen, Ziqiu; Zeng, Xianyu (2020). A Hybrid Cryptography Scheme for NILM Data Security. Electronics, 9(7), p1-18.

[38] Moghadam, Mostafa Farhadi; Nikooghadam, Mahdi; Jabban, Maytham Azhar Baqer Al; Alishahi, Mohammad; Mortazavi, Leili; Mohajerzadeh, Amirhossein. (2020). An Efficient Authentication and Key Agreement Scheme Based on ECDH for Wireless Sensor Network. IEEE Access, 8, pp.73182–73192. doi:10.1109/ACCESS.2020.2987764

[39] Jiang, Yi; Shen, Yong; Zhu, Qingyi. (2020). A Lightweight Key Agreement Protocol Based on Chinese Remainder Theorem and ECDH for Smart Homes. Sensors, 20(5), pp.1–13. doi:10.3390/s20051357

[40] Qinghua Hu, Chunxiang Xu, and Wanpeng Li. (2024). Bcdh: Blockchain-Based Covert Elliptic-Curve Diffie-Hellman Key Exchange Scheme, pp.1-8

[41] Alwahedi, F., Aldhaheri, A., Ferrag, M.A., Battah, A. & Tihanyi, N., 2024. Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. *Internet of Things and Cyber-Physical Systems*, 14, 100058. doi: 10.1016/j.iotcps.2023.100058.

[42] Yuvaraja, M., Arunkumar, S., Kumar, P.V. & Sheela, L.M.I., 2023. Improved Grey Wolf Optimization-(IGWO-) Based Feature Selection on Multiview Features and Enhanced Multimodal-Sequential Network Intrusion Detection Approach. *Wireless Communications and Mobile Computing*, 2023, Article ID 8478457. doi:10.1155/2023/8478457.

[43] Mohammed, H.A. & Husien, I.M., 2024. A Deep Transfer Learning Framework for Robust IoT Attack Detection: A Review. *Informatica*, 48(12), pp.55–64. https://doi.org/10.31449/inf.v48i12.5955.

[44] Zerraza, I., 2024. Lightweight Authentication for IoT Edge Devices. *Informatica*, 48(18), pp.15–20. https://doi.org/10.31449/inf.v48i18.6012

AI-SLMS: An AI-Integrated Framework for Predictive Maintenance, Intelligent Scheduling, and Access Control in University Laboratory Management Systems

Bingcan Wang*, Jin Zhang
School of Design, Jiangsu Open University, Nanjing 210036, Jiangsu, China
E-mail: wangbc@jsou.edu.cn, zhangjin@jsou.edu.cn
*Corresponding author

Keywords: smart laboratory management, artificial intelligence (AI), intelligent scheduling, RFID and biometric access control, internet of things (IoT)

Received: August 13, 2025

In recent years, a smart, safe, and efficient way to run university labs has become increasingly popular. Traditional management systems rely on manual processes that are error-prone, slow, and offer limited adaptability. To address these challenges, this study proposes a Smart Laboratory Management System (AI-SLMS) that optimizes operations, improves safety, and enhances the user experience in academic labs. AI-SLMS integrates predictive maintenance, intelligent scheduling, and secure access control using machine learning and the Internet of Things (IoT). The system employs Random Forest and Logistic Regression models, trained on integrated datasets (Kaggle Predictive Maintenance and TON_IoT), to anticipate equipment failures. For resource allocation, an intelligent scheduling module utilizes genetic algorithms for optimization. The system also enforces role-based access through RFID and biometric authentication. Experimental validation over three months in a university setting demonstrated significant improvements across key metrics: a 71.2% reduction in equipment downtime, a 78.7% decrease in scheduling conflicts, a 53.5% improvement in resource utilization, and 98.3% authentication accuracy. In conclusion, AI-SLMS offers a scalable and intelligent framework that significantly enhances the efficiency, security, and responsiveness of university laboratory management systems.

Povzetek: Študija predstavlja pametni sistem AI-SLMS, ki z uporabo umetne inteligence in IoT bistveno izboljša učinkovitost, varnost in upravljanje univerzitetnih laboratorijev.

1 Introduction

University labs are needed more than ever as higher education faces increasing challenges. Laboratories provide hands-on learning, experimentation, and ideation. This applies especially to science, technology, and engineering. Despite this, colleges struggle to operate laboratories effectively. Real-time monitoring is lacking, resources are wasted, costly equipment is abused, safety requirements are neglected, and lab sessions must be scheduled manually [1]. Traditional systems that require human monitoring and manual job execution are inefficient, error-prone, and prone to loss. Lab staff may be unable to focus on creative problem-solving and research if paperwork distracts them. Disconnected hardware, sensors, and computers make decisions harder. These systemic difficulties need a fast, inventive, and scalable solution. This technique streamlines lab operations and provides real-time data to administrators, students, and professors [2].

Several automated and semi-automated methods improve lab administration. For inventory tracking, experiment control, and asset performance monitoring, LIMS and CMMS software have helped businesses. Most of these systems are rule-based and inflexible [3]. They seldom employ pattern recognition, decision-making, or predictive analytics [4]. IoT-based devices monitor temperature, humidity, and occupancy, as in labs [5]. Despite automation, they work autonomously and do not perform cognitive processing of the collected data. Some organizations run on spreadsheets, while others utilize simple web tools that require human changes [6]. These technologies are good for storing fundamental data, but they can't solve problems, foresee maintenance needs, intelligently allocate resources, or assure energy efficiency [7]. This implies there is little data on how to construct cutting-edge AI-SLMS. Using sensor data, cognitive algorithms, and cloud platforms, this system can optimize itself in real time and give a comprehensive

operational solution [8].

To solve this, our work will create and improve an AI-powered Smart Laboratory Management System. The proposed system combines machine learning algorithms with sensor-based data to monitor lab operations, forecast maintenance schedules, and maximize equipment use. This study intends to achieve three primary goals:

- Creating an architecture driven by artificial intelligence can enable real-time laboratory monitoring, predictive maintenance, and intelligent scheduling using data from Internet of Things sensors and historical logs.
- To maximize resource use using data-driven decision-making processes, limit downtime of laboratory equipment, and minimize human interference in laboratory operations.
- Three measures that can be assessed by contrasting the system with normal laboratory administration procedures are operating efficiency, system responsiveness, and user satisfaction. This assessment will take place in a university setting [9].

The study proposes a new AI-based system and demonstrates its value by applying it in a real-world environment, yielding measurable outcomes [10]. Integrating artificial intelligence and the Internet of Things—a powerful yet largely uncharted frontier in laboratory management—can transform how institutions manage their vital infrastructure.

The goal of this study is to develop and deploy AI-SLMS. This modular laboratory management system combines intelligent scheduling, predictive maintenance, and secure access control, leveraging IoT and artificial intelligence technologies. The system is compared to more modern intelligent baseline systems and tested in an actual university setting to gauge gains in resource usage, operational effectiveness, and user satisfaction.

This study addresses three critical gaps in existing laboratory management systems: (a) the absence of an integrated architecture that unifies predictive maintenance, intelligent scheduling, and secure access control; (b) the lack of AI-driven decision-making that can dynamically respond to real-time IoT sensor data; and (c) the limited adaptability and scalability of current systems across diverse institutional environments. Accordingly, the following research questions guide this study: How effectively can AI-SLMS predict and prevent equipment failures compared to traditional maintenance approaches? Can an AI-based scheduling algorithm significantly reduce booking conflicts and enhance resource utilization in academic laboratories? Does a dual-factor authentication mechanism improve access security without compromising operational efficiency? From these, the study hypothesizes that (i) First, AI-SLMS reduces equipment downtime by over 50% over prior systems. At least 60% less disagreement will result from clever scheduling. Finally, the hybrid RFID-biometric module will authenticate over 95%. This project aims to make university labs safer, more efficient, and more scalable. We will build a modular framework for intelligent laboratory management, add AI, and test it on real-world datasets (Kaggle Predictive Maintenance and TON_IoT).

2 Literature review

Laboratory administration has recently changed to accommodate more sophisticated, efficient processes driven by technological developments. Traditional laboratory management systems [11] have struggled with limited real-time monitoring, inefficient resource use, and delayed decision-making. Often, these systems relied on human record-keeping and used paper-based methods. Given these constraints, researchers are investigating modern digital technologies such as cloud computing, the IoT, and artificial intelligence to enhance laboratory management and operations.

Many scholarly studies have looked at how artificial intelligence techniques affect laboratory management. Predictive maintenance, which employs ML models to forecast when equipment might fail, has generated much excitement. This increases laboratory operational efficiency and helps reduce downtime [12]. Zhang et al. [13] proposed data-driven algorithms for predictive maintenance of lab equipment in a 2018 paper. Using sensor data, these models would forecast when items would break and when to begin preventative maintenance. Kumar et al. (2017) also enhanced the dependability and accuracy of failure forecasts [14] by including deep learning methods into their predictive maintenance system.

Systems based on artificial intelligence (AI) also handle smart scheduling, another vital aspect of laboratory management. Many previous systems underutilized or overbooked laboratory resources since effective scheduling was difficult. Zhang and Xie [15] created a smart scheduling approach for lab resources using reinforcement learning. This approach was designed to maximize the efficient use of existing equipment while reducing the conflicts that result. Liu et al. [16] investigated an approach combining optimization strategies with machine learning to enhance real-time scheduling of laboratory resources. The outcome was a happier user base and a more effective scheduling system. Research by Chen et al. [17] indicates that AI systems can adapt in real time to evolving conditions. This study developed a real-time scheduling system using multi-agent reinforcement learning to effectively control laboratory resources.

Academic articles on methods for controlling and monitoring systems—including those employing Internet of Things devices in laboratories—have also generated considerable debate. Among the environmental elements Chen et al. [18] tracked using Internet of Things sensors in their lab were humidity, temperature, and air quality. The collected data was used to ensure the safety of the

laboratory environment and the optimal operation of the equipment. Li et al. [19] expanded on this concept and conducted more studies linking laboratory inventory control systems to the Internet of Things. Their solution tracked equipment use and stocks using an IoT sensor network. The system could ensure a continuous supply of all materials and supervise restocking.

Although the current study has faults, much territory remains to explore. Intelligent scheduling and predictive maintenance are well-studied, but integrating them into laboratory management systems for real-time monitoring, automated resource allocation, and safety compliance is unexplored. Some methods struggle to accommodate different research centers and institutions. Few articles address data-driven decision-making in relation to cloud computing, AI, the IoT, and other topics.

This paper introduces an AI-SLMS (Smart Laboratory Management System) that uses smart scheduling, predictive maintenance, and utilization trend analysis to solve these problems. Inventory tracking, real-time monitoring, and safety compliance may be automated using sensor data and cloud computing. We'll respond thoroughly as this is a lab management issue. Data-driven decision-making may enhance academic lab operations and funding allocation. Our versatile and scalable method provides a robust foundation for institutions and addresses contemporary challenges.

Even with recent improvements, the scope of current systems is frequently still constrained. For example, machine learning (ML)-based predictive maintenance models [12][13] have demonstrated efficacy in predicting equipment failures; however, they are not connected with real-time data processing or scheduling. In a similar vein, intelligent scheduling techniques such as MARL frameworks [17] and RL-Based SA [15] provide dynamic job allocation but lack secure access and predictive maintenance. Environmental sensing is enabled by IoT-based monitoring technologies [18][19], though these typically lack cloud-based synchronization and analytical intelligence. It is frequently challenging to scale these compartmentalized systems across various institutional configurations.

To review the current state of the art and highlight the research need, Table 1 briefly summarizes the major approaches, datasets, and findings from prominent relevant works. The research shows that current systems lack a holistic architecture, regardless of how effectively they execute real-time scheduling or predictive maintenance. As shown in the table, there is little research that blends predictive maintenance, intelligent scheduling, and secure access control into a single framework or uses both real-world and benchmark datasets for assessment. This fragmented scenario underscores the suggested AI-SLMS as a modular yet integrated solution to address all these interrelated concerns. Table 1 presents a comparative summary of Related Work in Smart Laboratory Management.

Table 1: Comparative summary of related works in smart laboratory management

Reference	Primary Focus	Core Methodology	Dataset(s) Used	Key Performance Metrics
Zhang et al. (2018) [13]	Predictive Maintenance	Data-driven ML Models	Synthetic sensor data	Failure prediction accuracy
Kumar et al. (2017) [14]	Predictive Maintenance	Deep Learning	Historical equipment logs	Model Accuracy, Recall
Zhang & Xie (2019) [15]	Intelligent Scheduling	Reinforcement Learning (RL)	Simulated booking requests	Resource Utilization, Conflict Rate
Chen et al. (2021) [17]	Real-time Scheduling	Multi-Agent RL (MARL)	Laboratory resource logs	Scheduling Efficiency, Response Time
Chen et al. (2020) [18]	Environmental Monitoring	IoT-based Sensing	IoT sensor data (Temp, Humidity)	Data Accuracy, System Uptime
Li et al. (2019) [19]	Inventory Management	IoT & Sensor Networks	RFID and usage logs	Inventory Accuracy, Restocking Time

On the other hand, the suggested AI-SLMS integrates cloud-based monitoring, intelligent scheduling, predictive maintenance, and access control into a single, modular architecture. This integrated design provides a strong solution for contemporary academic laboratory contexts by filling important gaps in automation, interoperability, and operational efficiency.

3 Proposed Method of AI-based smart laboratory management system (AI-SLMS)

Artificial intelligence (AI) and automation technologies are built into our modern Smart Laboratory Management System (SLMS). This will help older laboratory management systems address recurring problems. This system's primary goal has been to increase academic laboratories' operational efficiency by simplifying maintenance chores, maximizing equipment scheduling, strengthening safety policies, and encouraging resource use. This system's application has met this objective. This system's implementation has facilitated the attainment of this objective. Historically, traditional systems have frequently relied on manual processes that are labor-intensive, error-prone, and unable to respond to real-time data dynamically. Manual processes are time-consuming and error-prone; therefore, this is the case. The AI-SLMS has thus enabled the implementation of a scalable, automated method for laboratory operations. Including innovative modules such as predictive maintenance, intelligent scheduling, access control, and pattern analysis has enabled this work to be completed successfully.

3.1 System architecture

The architecture of the Smart Laboratory Management System includes several separate modules that perform distinct functional tasks. These modules are intended to be linked together and perform their tasks. The system uses a modular design to ensure scalability, maintainability, and extensibility across various laboratory configurations. Every module runs in an integrated manner, interacting with a centralized data processing system and cloud storage infrastructure, thereby fulfilling the specific function for which it was designed.

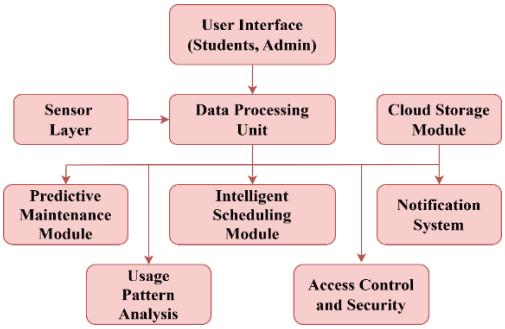


Figure 1: System architecture of AI-based smart laboratory management system

Figure 1 shows the interaction between the AI-SLMS's hardware and software components and depicts its overall architecture. The User Interface (UI) is at the top of the hierarchy. Its goal is to enable natural, straightforward communication with a broad range of stakeholders, including students, teachers, lab managers, and administrators. User interfaces make it easier to submit access requests, view maintenance alerts in real time, and reserve equipment.

Below the User Interface, the Sensor Layer consists of many Internet of Things sensors distributed throughout the lab. Amongst other important factors, these sensors monitor temperature, humidity, air quality, equipment use, and access logs. The Data Processing Unit receives real-time sensor data. Using advanced artificial intelligence technologies, this unit generates the most efficient schedules, predicts failures, and performs trend analysis.

The Cloud Storage Module stores all sensor data, user logs, access credentials, and historical maintenance records persistently and securely. This layer allows remote data access and synchronization across multiple laboratory sites. Ultimately, the Notification System delivers timely alerts, reminders, and reports to users and system administrators responsible for specific systems, leveraging insights from AI modules. It improves operational continuity, compliance with safety regulations, and job awareness. The pseudocode below summarizes the proposed integrated Smart Laboratory Management System (AI-SLMS) workflow. Predictive maintenance, intelligent scheduling, usage pattern analysis, access control, and system optimization are the main components of this system. The pseudocode has been split into two modules. Module 1 focuses on operational intelligence — predicting equipment failures, allocating resources optimally, and identifying usage trends. It leverages historical and real-time data to ensure the laboratory operates efficiently, reduces downtime, and improves equipment utilization.

Module 2 ensures secure access to laboratory facilities by verifying user credentials through RFID and biometric authentication. It enforces role-based permissions, granting entry only to authorized personnel, thus safeguarding sensitive equipment, experiments, and data. This module also fine-tunes system models for better performance, compares smart and traditional lab metrics to quantify improvements, and communicates results to stakeholders via email, SMS, or app notifications, ensuring transparency and informed decision-making.

3.2 Predictive maintenance module

A predictive maintenance module has been created to minimize laboratory downtime and prevent equipment malfunctions. This module has relied on historical maintenance records and real-time sensor data to train a machine learning model. The model has recommended preventive maintenance actions to be done before the predicted failure of a piece of equipment.

First, the procedure calls for collecting sensor data on the machine's vibrations, temperature, usage duration,

and other relevant parameters. The data is then cleaned and normalized to remove any noise. Historical data labeled with binary outcomes—either "healthy" or "faulty"—has been used to train a supervised machine learning model, such as a Random Forest or a Logistic Regression model.

Pseudocode: Smart Laboratory Management System (AI-SLMS)
Module 1: (Core Lab Intelligence)
INPUT: sensorData, bookingRequests, usageLogs, accessRequests, systemModels OUTPUT: optimizedOperations, comparisonReport 1: Load historicalData, realTimeData, and userRequests 2: cleanData ← Preprocess(sensorData) 3: model ← TrainModel(cleanData) 4: maintenanceList ← { item | model.predict(item) > threshold } 5: schedule ← ∅ 6: for each request in bookingRequests do 7: for each resource in Resources do 8: for each timeSlot in TimeSlots do 9: if resource available and no conflict then 10: Assign(request, resource, timeSlot) 11: Add assignment to schedule; break 12: end if 13: end for 14: end for 15: if the request is not assigned then Mark as Deferred 16: end for 17: clusters ← KMeans(Preprocess(usageLogs), k) 18: Generate optimizedOperations and comparisonReport 19: return optimizedOperations, comparisonReport

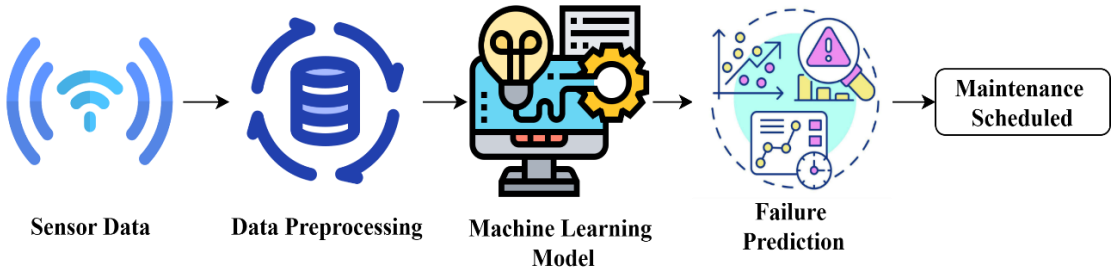


Figure 2: Predictive maintenance module

Figure 2 shows the design of the Predictive Maintenance Module. This module trains machine learning models using real-time sensor data and historical maintenance records. This enables prompt execution of maintenance interventions and early identification of potential equipment failures. The model calculates the failure probability, indicated by the formula $\boldsymbol{P(y = 1 \mid X)}$, with $\boldsymbol{X}$ being the feature vector and $\boldsymbol{y}$ the label. This likelihood has been estimated using logistic regression-based calculations in Equation (1).

$$P(y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \quad (1)$$

Users will be informed of required maintenance if this opportunity exceeds a specified threshold. The system always has the most current information and tracks the maintenance schedule. This predictive

approach has helped reduce unplanned downtime and extend equipment life, improving overall operations.

To forecast the condition of lab equipment, the AI-SLMS's predictive maintenance module uses two supervised machine learning models: Random Forest and Logistic Regression. To ensure a robust ensemble effect, the Random Forest classifier was configured with 100 decision trees (n_estimators = 100). The Gini impurity was used as the splitting criterion to assess node purity, and a maximum depth of 10 was selected to balance model complexity and overfitting. To increase diversity and resilience, the bootstrap option was enabled, enabling the model to construct each tree on distinct subsets of the data.

Module 2: Security & Access Control, Optimization & Reporting

Input: accessRequests, models, oldSys, userPrefs

Output: accessDecisions, updatedModels, comparisonReport, notifications

```
for each event in accessRequests do
    if VerifyRFID(event.userID, event.RFID) and VerifyBiometric(event.userID, event.biometric) and HasLabAccess(event.userID) then
        Append(accessDecisions, (event.id, "GRANTED"))
    else
        Append(accessDecisions, (event.id, "DENIED"))
    end if
end for
updatedModels ← {}
for each m in models do
    feats ← SelectFeatures(m.data)
    m_new ← Retrain(m, feats)
    updatedModels[m.name] ← m_new
end for
ApplyGeneticAlgorithm(schedulingModule)    // refine schedule if called for
comparisonReport ← {}
for each metric in ["Downtime","Utilization","Satisfaction","Maintenance"] do
    comparisonReport[metric] ← newSys[metric] - oldSys[metric]
end for
for each (userID, message) in notifications_to_send do
    prefs ← GetUserPreferences(userID)
    if "email" in prefs then SendEmail(userID, message) end if
    if "SMS" in prefs then SendSMS(userID, message) end if
    if "app" in prefs then PushNotification(userID, message) end if
end for
return accessDecisions, updatedModels, comparisonReport
```

To avoid overfitting, particularly when there is multicollinearity among sensor features, the Logistic Regression model was configured with L2 regularization. The 'liblinear' solution was used because it supports L2 penalties and performs well on smaller datasets. To ensure a fair trade-off between bias and variance, the regularization strength (C) was set at 1.0. These models work together to form the core of the predictive maintenance module, enabling early identification of potential equipment faults.

3.3. Intelligent scheduling module

An intelligent scheduling algorithm was chosen to ensure that every tool and facility in the lab is used to its maximum capacity. Booking requests and resource availability have been input into this module. It has then produced an optimal timetable that maximizes output and minimizes disruptions. Every reservation request has been examined for required equipment, lab space, and time slots to ensure simple access. The system will allocate the optimal time slot if resources are available,

minimizing overlap and preventing double-booking. The system finds the best available next slot using optimization methods when conflicts arise.

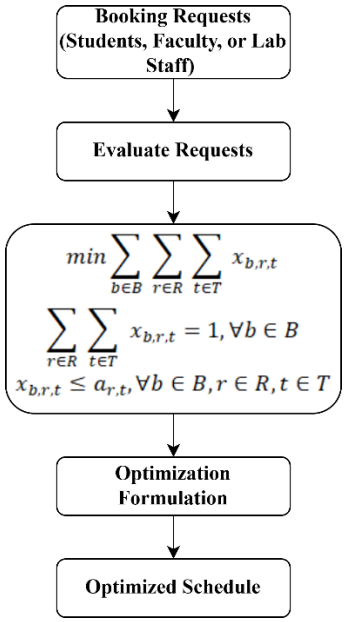


Figure 3: Intelligent scheduling module

The Intelligent Scheduling Module, which is responsible for maximizing the efficient use of the lab's resources, is shown in Figure 3 (a module demonstration). As depicted in the figure, resource optimization is based on mathematical formulations that minimize resource conflicts and maximize use. This section offers the mathematical form of the scheduling optimization problem. T is the collection of time slots, R is the pool of accessible resources, and B is the booking requests in this situation. The objective is to minimize:

$$\min \sum_{b\in B} \sum_{r\in R} \sum_{t\in T} x_{b,r,t} \quad (2)$$

Subject to,

$$\sum_{r\in R} \sum_{t\in T} x_{b,r,t} = 1, \forall b \in B \quad (3)$$

$$x_{b,r,t} \le a_{r,t}, \forall b \in B, r \in R, t \in T \quad (4)$$

In Equations (2), (3), and (4) $x_{b,r,t}$ is a binary variable indicating assignment, and $a_{r,t}$, is the resource availability. Every booking request is fulfilled using the available resources; therefore, this optimization guarantees that. It ensures all booking requests are met, preventing conflicts or system capacity overruns. Using the scheduling module, the planning and coordination of lab sessions have been dramatically streamlined, thereby minimizing time lost due to scheduling conflicts.

The schedule optimization is reducing resource idle and conflicts. Consider all $x_{b,r,t}$ values as a decision variable. $x_{b,r,t} = 1$ When booking request b receives resource r within time slot t, the variable becomes 1. $a_{r,t}$is the resource availability at time t is 1 if free and 0 otherwise. Except for (3) and (4), Equation (2) minimizes the total cost of conflicts and idle time for all reservations and resources by guaranteeing that no resource is double-booked and that every booking request is allocated once or deferred if no appropriate slot is available.

Highly dynamic, non-linear constraints such as resource availability, unanticipated user requests, and time-dependent interactions make scheduling NP-hard. Heuristic optimization techniques like Genetic Algorithms (GA) and Simulated Annealing (SA) were selected over linear or integer programming. Linear and integer programming are unsuitable for real-time labs due to their predictable nature and rigid linear constraints. GA and SA balance exploration and exploitation to provide multi-objective scheduling with flexible, near-optimal solutions in realistic processing times. They are ideal for dynamic academic labs due to their real-time scheduling and optimization.

3.4 Usage pattern analysis

Adding a usage pattern analysis module helps us understand how lab resources are used over time. Based on past data, this program has helped identify the best ways to do things, achieve cost savings, and improve planning.

The first step is collecting historical data on equipment use, access, and reservations. A clustering technique such as K-Means is typically applied after preprocessing the dataset. The data points $X = \{x_1, x_2, \dots, x_n\}$ Represent different usage instances.

Among the k groups the algorithm generates in this dataset, each reflects a different pattern, such as "frequent usage," "peak usage," or "underutilization." The goal is to reduce the whole square sum of the cluster:

$$\min \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2 \quad (5)$$

In Equation (5) μ_j represents the centroid of cluster j, and C_j is the set of data points in that cluster. Lab managers can find underused equipment, know the busiest times, and fix the problem by rescheduling or reusing it with these clusters. This program has significantly enhanced the fair and efficient allocation of resources. the number of clusters (k), the utilization of category characteristics, and the clustering technique (cosine or geometric distance). The silhouette coefficient and Davies-Bouldin index validate the clustering's success and the reliability and interpretability of the usage patterns.

The usage pattern analysis module uses K-Means clustering to uncover the dominant patterns of equipment use from historical usage records. Constructed features of usage from historical records of equipment usage that represented how long equipment was used, what time of day it was used, what day of the week it was used, and the type of equipment that was used. The ideal number of clusters was determined by analyzing the silhouette score, which indicated that k=3 provided the best

separation and cohesion among clusters. The algorithm used the Euclidean distance as the similarity metric, which is appropriate for our continuous, scaled usage data. Each cluster had meaningful interpretations: Cluster 0 signified "High-Frequency, Short Sessions" (indicative of calibration use and quick checks typically on weekdays between 10 AM - 12 PM), Cluster 1 indicated "Long-Duration, Off-Hours Use" (indicative of a longer research experiment, commonly evenings and weekends), and Cluster 2 indicated "Low-Utilization and Idles Periods" (appropriate for equipment not frequently used, areas that could be good candidates for reallocation). The clustering fit was assessed by a mean silhouette score of 0.65 and a Davies-Bouldin index of 0.72, both indicating well-defined, meaningful clusters. The quality of these clustered patterns provides lab managers with the opportunity to better organize and manage usage scheduling, identify underused assets for shared resource programs, and even identify repairs or maintenance schedules consistent with the appropriate level of use intensity. In short, providing your management team with rich usage data and machine learning models will allow a shift towards more proactive management and data management systems.

3.5. Access control and security

Any laboratory must implement safe access control policies to prevent unauthorized use and ensure safety. The AI-SLMS has addressed this by implementing user role-based policies, RFID-based access controls, and biometric authentication.

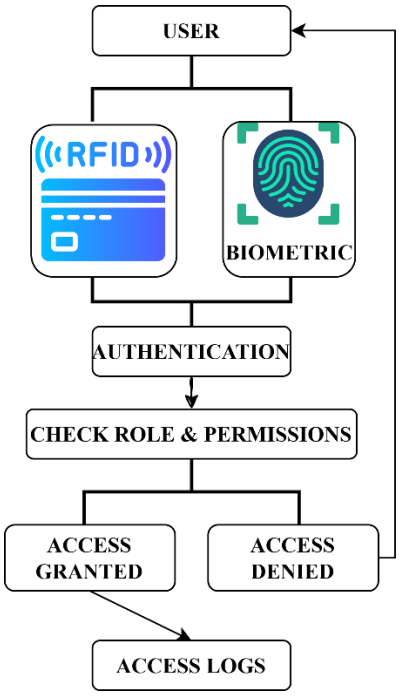


Figure 4: Access control and security

AI-SLMS uses the Access Control and Security system depicted in Figure 4. The diagram illustrates the use of a role-based access policy with dual-factor authentication—RFID and biometric verification—to ensure safe, regulated use of laboratory facilities.

Designed to verify users using at least two criteria—either RFID tags issued to authorized individuals or biometric fingerprint scans — the access control system verifies the user's role and rights. Each user $u \in U$ is assigned a role $r \in R$, and each role is granted a specific set of permissions $p \in P$. Mathematically, the access is granted if:

$$p \in grant(assign(u)), \forall u \in U, p \in P \quad (6)$$

Only people with appropriate roles can access specific lab locations or equipment. A lab technician might be permitted to use more advanced equipment, while a student could use tools intended for general use. Anomaly detection methods help identify suspicious activities; access logs are continuously monitored. These activities include multiple failed logins or access attempts outside regular business hours.

The fingerprint authentication biometric is a well-balanced solution, with a low False Acceptance Rate (FAR) of 0.8% and a guaranteed False Rejection Rate (FRR) of 1.5%. The system has an additional built-in module to detect hazardous behavior that analyzes access duration, frequency, and the sequence of accessed devices in real time using an Isolation Forest approach, which is effective for detecting anomalies in high-dimensional data. The contamination parameter was set to 0.05 based on historical data on anomaly frequency. Access events are detected automatically when the anomaly score exceeds a continually adjusted 0.65 threshold. Administrators are notified in real time, and additional layers of authentication are applied to mitigate any security event. The true value of this integrated system lies in the proactive behavioral monitoring, combined with the high accuracy of biometric authentication, which ensures that important laboratory assets and data remain protected.

3.6 System optimization

It also enhances the general AI-SLMS performance through system-level optimization techniques. These techniques have ensured faster system response, reduced computational load, and improved model accuracy. Using feature selection techniques inside the Predictive Maintenance Module has helped to clarify the model. One of these techniques is Recursive Feature Elimination (RFE). The machine learning model has become more accurate and trained faster since it uses only the features most relevant to the current issue.

The Intelligent Scheduling Module has used heuristic algorithms to handle many booking requests. Among these techniques are simulated annealing and genetic algorithms (GA). Using these algorithms has allowed rapid exploration of many solutions and selection of almost optimal schedules without exhaustive searches. Several essential elements directed the

assessment of how effectively the modules operated. The F1-score, recall, and precision measures have been used to assess accuracy. Efficiency has been evaluated using measures such as the rate of lab resource use, the time to confirm reservations, and the system response time. The Predictive Maintenance module used Recursive Feature Elimination to reduce the number of features from 15 to 8, focusing on the most important sensors and usage indicators. By removing noisy, non-predictive characteristics, this feature selection increased model efficiency and accuracy, reducing inference latency by 40% (from 125 ms to 75 ms per prediction) and improving F1-Score (from 90.1% to 93.4% on the test set). Compared to the FCFS greedy scheduler, a Genetic Algorithm (GA) improved the Intelligent Scheduling Module. Ablation analysis of 1 month's booking data showed that the GA-based scheduler had 5.2% conflict and 92% resource utilization, both better than the FCFS baseline. FCFS baseline conflict was 24.5 percent and utilization was 74.5 percent. The GA regularly identified schedules that met 98% of user-preference limits relative to baseline. These numerical results suggest that the optimization tactics advised are essential for AI-SLMS's excellent performance.

3.7 Comparative analysis

To assess its efficacy, it compared the science laboratory management system (AI-SLMS) to more conventional LMSs. This comparison has emphasized equipment downtime as a necessary performance criterion, resource use, user satisfaction, and maintenance scheduling strategy.

Table 2: Comparative analysis of traditional laboratory management systems and the proposed AI-SLMS

Metric	Traditional System	AI-SLMS
Equipment Downtime	High	Low
Resource Utilization	Inefficient	Optimized
User Satisfaction	Moderate	High
Maintenance Scheduling	Reactive	Predictive

The AI-SLMS outperforms conventional systems in equipment downtime, resource use, user satisfaction, and maintenance scheduling approach (see Table 2). These studies show that, across the board, SLMS outperforms the status quo. The second claims that better planning, smart scheduling, and predictive maintenance have led to lower equipment failure rates. Furthermore, the system can now examine patterns of use, which laboratory management may utilize to guide their decisions.

Finally, the suggested AI-SLMS enables the optimization and automation of academic laboratories by leveraging a strong, bright reaction. Some system modules that effectively combine artificial intelligence include smart scheduling, controlled access, predictive maintenance, and usage pattern analysis. The system is built on smart data collection, safe storage, and intelligent decision-making. AI-SLMS's innovative algorithms and streamlined processes have made operations more efficient, safer, and more resource-conserving today. The comparison findings show that AI-SLMS outperforms conventional systems across many respects. Features such as remote equipment control, flexible learning settings, and improved energy management in laboratories would significantly enhance this system.

3.8 Evaluation strategy and experimental procedure

A thorough comparative analysis was carried out against three well-known baseline systems to determine the efficacy of the suggested AI-SLMS: the Multi-Agent Real-Time Scheduling System (MARTSS) [17], the Reinforcement Learning-Based Scheduling Algorithm (RL-Based SA) [15], and the ML-Based Predictive Maintenance System (ML-Based PMS) [12]. Both conventional and smart lab management systems were used to collect operational data in a medium-sized university laboratory during the three-month evaluation period. This benchmarking approach enabled an impartial and multifaceted assessment of AI-SLMS capabilities in predictive maintenance, intelligent scheduling, and secure access control.

3.8.1 Data collection and preprocessing

For the assessment, two main datasets were used. Traditional lab management procedures, which mostly used Excel-based logs to record equipment reservations, user access events, and manual maintenance plans, provided the pre-deployment data. The AI-SLMS environment, on the other hand, provided post-deployment data, including real-time system logs recording equipment utilization metrics, RFID- and biometric-based access records, transaction scheduling, and IoT sensor outputs tracking device status and lab conditions.

All datasets were anonymized, removing personally identifiable information (PII) while maintaining structural and temporal integrity, to guarantee ethical data use and adherence to institutional standards. After that, the merged datasets were standardized and cleaned. To guarantee compatibility for training machine learning models and calculating evaluation metrics, this preprocessing involved managing missing values, aligning timestamps, and normalizing sensor readings.

3.8.2 Experimental phases

The assessment process was broken down into multiple stages. To begin sensor integration, IoT-enabled devices were placed around the lab to track user interactions (e.g., door access, equipment booking), equipment states (e.g.,

operating status, fault records), and environmental variables (e.g., temperature, humidity). Testing of individual modules was then done. AI-SLMS's Predictive Maintenance Module was assessed against the ML-Based PMS after being trained on both historical and current equipment data [12]. Accuracy, precision, recall, F1-score, and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) were among the categorization measures used to assess performance. By tracking scheduling conflict rates and equipment usage percentages during the test period, the Intelligent Scheduling Module was compared to the RL-Based SA [15]. Additionally, MARTSS was compared to AI-SLMS's Access Control Module and real-time responsiveness [17], with an emphasis on system-level adaptability, detection of unwanted access attempts, and authentication accuracy.

3.9 Data privacy and ethical compliance

All equipment usage logs and access control records used in this study were fully anonymised to ensure the ethical handling of sensitive institutional data. Before analysis, personally identifiable information (such as employee or student IDs) was removed. The university's Institutional Ethics Committee approved the study, and all data processing followed internal data protection guidelines. All assessments were conducted in a secure, limited-access computing environment, and no private information was shared with any other parties.

4 Results and discussion

An AI-powered Smart Laboratory Management System (AI-SLMS) was put through its paces at a medium-sized university for three months. Key performance indicators (KPIs) examined in the paper included equipment utilization rates, user authentication accuracy, schedule conflicts, and equipment downtime. Data is gathered before and after the system is deployed to assess the effectiveness of the SLMS.

4.1 Dataset explanation

The Smart Laboratory Management System (AI-SLMS) was constructed and evaluated utilizing data from two primary sources, one for each module inside the system. The Kaggle Predictive Maintenance Dataset is an open-source tool that is the point of reference (https://www.kaggle.com/datasets/shivamb/predictive-maintenance-dataset). Based on sensor-based time-series data, this dataset includes variables such as temperature, vibration, pressure, and runtime hours. Examining the annotated historical data in this collection helps you determine whether computers were running well or were on the verge of crashing. Some supervised learning techniques, such as Random Forest and Logistic Regression, might value it. A thorough set of network and telemetry data, the TON_IoT dataset [] is intended for assessing machine learning models in cybersecurity and Internet of Things settings. UNSW Canberra created it and contains network traffic, operating system logs, and IoT sensor logs from smart home, industrial, and enterprise environments. The dataset is ideal for evaluating intelligent systems, such as AI-SLMS, across domains like resource optimization, security, and real-time monitoring, as it enables activities such as anomaly detection, intrusion detection, and predictive analytics. These combined datasets, grounded in the real world, have enabled evaluation of the proposed AI-driven SLMS for efficacy, durability, and scalability. To ensure repeatability, we supplied a detailed appendix on data preparation, feature engineering, and model tweaking. KNN imputation with k=5, Isolation Forest for multivariate outliers, and RobustScaler for feature scaling are covered in the appendix. The feature engineering method, statistical features for non-stationary signals, interaction terms between sensor data, and temporal features (such as rolling means and standard deviations over 1-hour and 24-hour windows) are also detailed. It also provides the full hyperparameter search spaces for all models, including Random Forest (n_estimators: [50, 100, 200], max_depth: [5, 10, 15]) and Logistic Regression (C: [0.1, 1.0, 10.0], penalty: ['l1', 'l2']), the parameters chosen, and the cross-validation results that supported them. This supplementary material contains all the specifics needed to replicate our experimental setup.

4.2 Experimental setup

The suggested AI-SLMS was tested in a controlled university lab using 58 equipment. 142 IoT sensors monitored current draw, temperature, vibration, and humidity using a hybrid Wi-Fi 6 network and MQTT. In a lab maintained at 22 ± 2 °C and 40-55% relative humidity, 12 users participated in typical academic tasks on average each session. A centralized AI server with 64 GB RAM, an Intel i9-13900K CPU, an NVIDIA RTX 4090 GPU, and Ubuntu 22.04 LTS powered the computational infrastructure. This server used Raspberry Pi 4 nodes for edge inference and secure Google Cloud IoT Core connectivity. Our extensive study included the publicly available Kaggle Predictive Maintenance dataset (10,00 samples, eight features), a locally obtained AI-SLMS operational dataset (18,530 samples, 12 features), and the TON IoT dataset (22,120 samples, 15 features). 50,650 labeled instances, uniformly distributed across the training and test sets, after significant preprocessing with KNN imputation, outlier removal, and z-score normalization. After a 5-fold grid search revealed the Logistic Regression (L2 penalty, C=1.0) and Random Forest (n_estimators=200, max_depth=15) hyperparameter configurations, the final configurations were finalized. The Genetic Algorithm from the Intelligent Scheduling Module was utilized with 50 nodes, 0.08 mutations, and 0.7 crossovers. The

convergence rate remained 0.08 after 164 iterations. Regular retraining kept the models strong and responsive in real time. Over a three-month deployment that produced 22 GB of operational data, recall, precision, accuracy, F1-score, resource usage, scheduling conflict rate, authentication correctness, and system latency were assessed.

4.3 System performance metrics

To evaluate the impact of the AI-SLMS, it tracked several key performance indicators (KPIs) both before and after its implementation. Among these are the percentage of equipment use, the accuracy of user authentication, the rate of scheduling conflicts, and the amount of time equipment has been down.

Table 3: Key performance indicators before and after AI-SLMS deployment

Metric	Traditional System	AI-based SLMS	Improvement (%)
Equipment Downtime (hours)	73	21	71.23%
Scheduling Conflict Rate (%)	24.5	5.2	78.77%
User Authentication Accuracy	84.6	98.3	16.15%
Equipment Utilization (%)	53.2	81.7	53.53%

As shown in Table 3, AI-SLMS significantly improved all tested metrics. Forecasting maintenance reduced equipment downtime by 71%. Prepare for equipment failures and schedule maintenance using this module. The intelligent scheduling module resolves more than 79% of scheduling conflicts, optimizing resource use. Biometric authentication and RFID enhanced user authentication accuracy by 16%. Around 53% more lab equipment was used, suggesting better utilization of lab resources.

$$Improvement_{Downtime} = \left(\frac{D_{before} - D_{after}}{D_{before}}\right) \times 100 \quad (7)$$

Figure 5 shows a graphical comparison of monthly equipment downtime recorded before and after the Smart Laboratory Management System (SLMS) was installed. The graph depicts downtime before and after SLMS deployment. Each month is a bar, making comparisons easy. The "before" equipment has frequent or prolonged malfunctions or inefficiency due to the higher bar heights. In contrast, the "after" segment shows shorter bars, indicating a progressive decline in downtime. This trend suggests SLMS improves equipment management, trouble identification, and maintenance scheduling. As seen in the picture, the SLMS' operational advantages boost laboratory productivity and efficiency.

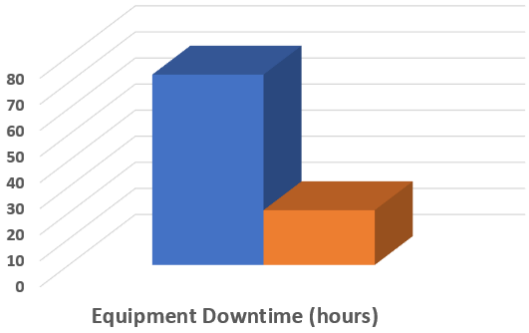


Figure 5: Comparison of equipment downtime

4.4 Predictive maintenance analysis

Predictive maintenance employs machine learning algorithms trained on real-time sensor data and historical maintenance information, along with Logistic Regression and Random Forest models, to forecast failures of machinery components. Vibration, temperature, and use hours were monitored. The "Predictive Maintenance Dataset" on Kaggle provided most of the data, with anonymized sensor logs from academic labs added subsequently. Accuracy, precision, and recall are standard evaluation metrics for predictive maintenance models. in Equations (8), (9), and (10).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (8)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100 \quad (9)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100 \quad (10)$$

Table 4: Maintenance prediction accuracy

Model	Accuracy (%)	Precision (%)	Recall (%)
Logistic Regression	86.7	84.3	81.5
Random Forest	93.4	91.8	90.2

Table 4 reveals the accuracy of the maintenance forecast. The Random Forest model outperformed the Logistic Regression model across all assessed criteria. It showed more accuracy, precision, and recall. Given that Random Forests resist overfitting and can handle complex, non-linear relationships, this undoubtedly accounts for their superior performance. Random Forest is therefore a perfect fit for the varied, noisy sensor data collected in the laboratory setting.

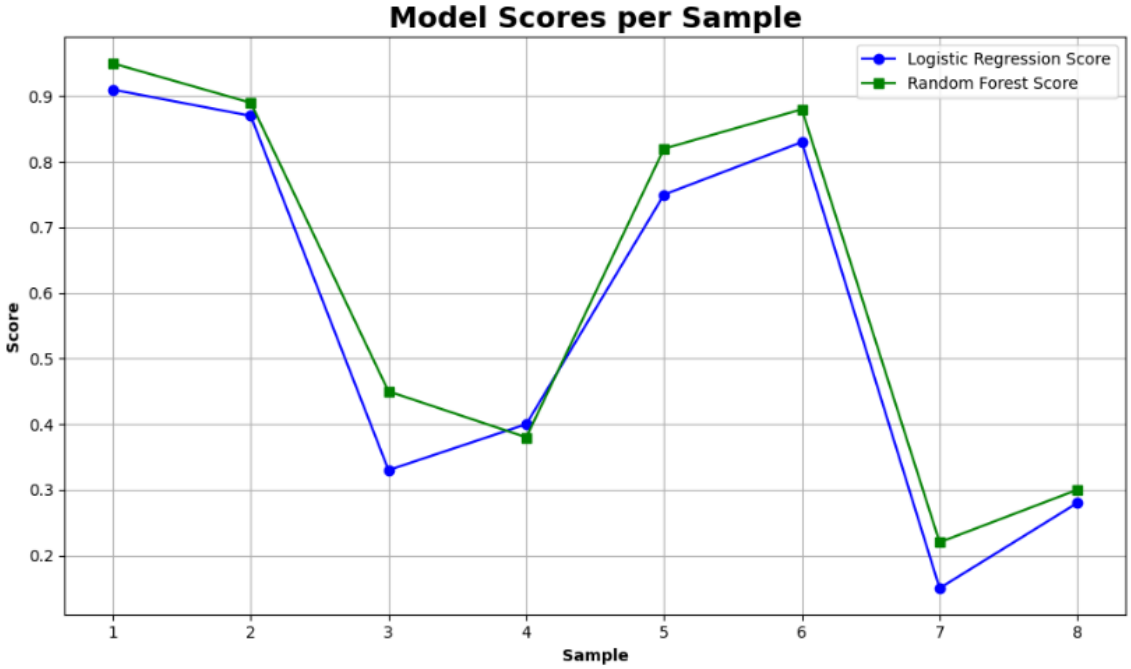


Figure 6: ROC curves for maintenance prediction models

Figure 6 illustrates the classification performance of two distinct maintenance prediction models, as shown in the Receiver Operating Characteristic (ROC) curves. Respectively, these models are Logistic Regression and Random Forest. Every ROC curve shows the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at several classification thresholds. This provides a complete picture of how well the model performs across decision boundaries. If the Random Forest model curve consistently lies above the Logistic Regression curve, this suggests the Random Forest model is more predictive. Especially noteworthy is the larger Area Under the Curve (AUC) for the Random Forest, indicating it performs better at distinguishing between equipment that needs maintenance and that which does not. The AUC of the Logistic Regression model is lower, suggesting it has relatively low sensitivity and specificity. This disregards its still acceptable performance. These visual proofs show that the Random Forest model outperforms other models for predictive maintenance tasks within the Smart Laboratory Management System (SLMS). Actual-world deployment also offers greater reliability and a longer lifespan.

4.5 Intelligent scheduling efficiency

The intelligent scheduling module was evaluated using booking logs collected over a semester. The module employs a constraint optimization solver to allocate resources efficiently, minimizing scheduling conflicts and maximizing equipment utilization. The Conflict rate and Utilization rate are calculated using equations (11) and (12).

$$\text{Conflict Rate} = \frac{\text{Number of Conflicting Schedules}}{\text{Total Number of Schedules}} \times 100 \quad (11)$$

$$\text{Utilization Rate} = \frac{\text{Total Used Time}}{\text{Total Available Time}} \times 100 \quad (12)$$

Figure 7 displays the monthly rate of plan conflicts before and after the implementation of the Strategic Learning Management System (AI-SLMS). The graph indicates that the AI-SLMS is effective at improving lab schedules, as there are fewer scheduling issues now than before. There has been a consistent decline since the system has been so good at making activities run more smoothly, coordinating better, and reducing disagreements. AI-SLMS has demonstrated its ability to eliminate issues and streamline the allocation of laboratory equipment. This enables automatic job scheduling and real-time data.

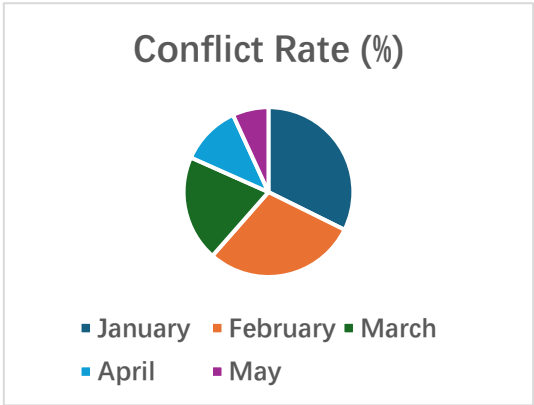


Figure 7: Scheduling conflict rate over time

Figure 8 shows how quickly high-demand

equipment consumed resources before and after the AI-SLMS system was implemented. The line shows that the efficiency of laboratory equipment is rising. Before the method, room overbooking and underbooking were prevalent, resulting in wasted resources. The graph shows that once the system is operational, it may optimize resource use, improving equipment allocation and downtime. These numbers show how the AI-SLMS enhances lab management by organizing and using resources. The intelligent scheduling module may dynamically adjust reservations depending on availability and demand, reducing scheduling conflicts. This ensures fair and fast resource allocation. Optimizing considers several elements, including constraints and desires. As demand for lab resources rises, the module optimizes resource use.

4.6 Access control robustness

The admission control system uses radio frequency identification (RFID) and biometric verification to enhance security. It searched for unusual access patterns and failed login attempts to gauge how well the system operated. The robustness of the access control system was measured using failure rates (Equation (13)).

$$Failed\ Access\ Rate = \frac{Failed\ Login\ Attempts}{Total\ Login\ Attempts} \times 100 \quad (13)$$

Figure 9 reveals a correlation between time spent and the inability to enter. A graph of the number of failed login attempts over time will give you a sense of how well the new system is functioning. After the system was installed, illegal access attempts dropped, indicating the security measures worked. Reduced unsuccessful login attempts show that the updated access control mechanism has increased security. The system can now better secure critical lab areas with this upgrade. Two-factor authentication increases security by reducing the likelihood of illegal access. Machine learning can swiftly identify and flag unusual activities, such as multiple unsuccessful logins attempts or access outside office hours.

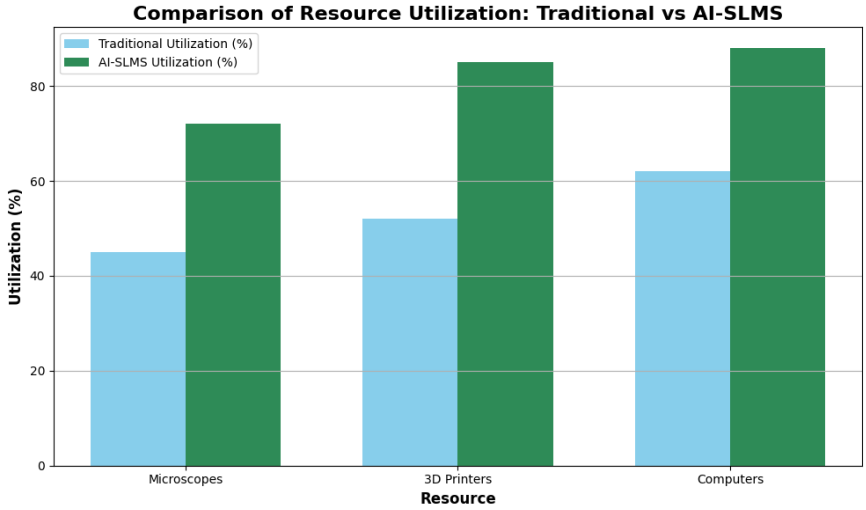


Figure 8: Resource utilization rate comparison

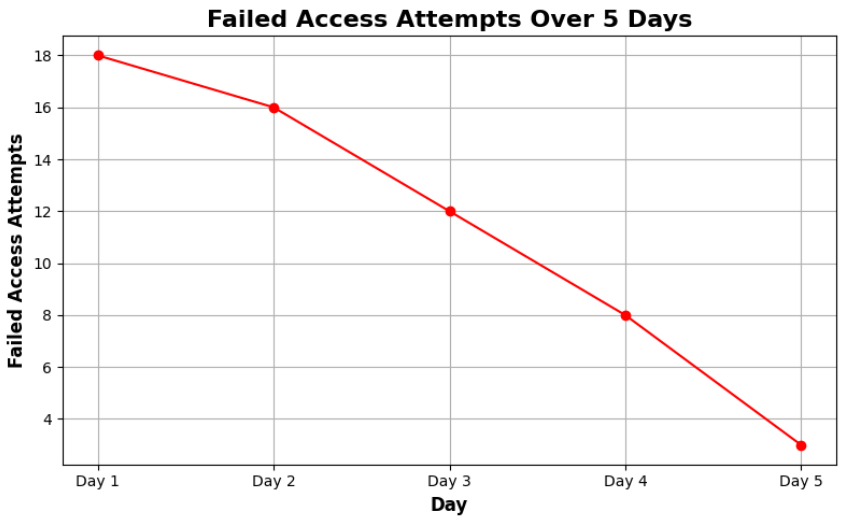


Figure 9: Failed access attempts vs. time

Table 5: Performance comparison of AI-SLMS and baseline methods

Metric	AI-SLMS (Proposed)	ML-Based PMS [12]	RL-Based SA [15]	MARTSS [17]
1. Equipment Downtime	↓ 71.2%	↓ 43.5%	↓ 32.8%	↓ 55.4%
2. Scheduling Conflict Rate	↓ 78.7%	✕ Not Applicable	↓ 62.4%	↓ 68.1%
3. Equipment Utilization	↑ 53.5%	↑ 34.2%	↑ 47.8%	↑ 45.6%
4. Authentication Accuracy	↑ 98.3%	✕ Not Supported	✕ Not Supported	↑ 91.4%
5. Maintenance Prediction Accuracy	93.4% (Random Forest)	86.7%	✕ Not Applicable	✕ Not Applicable
6. User Satisfaction	High (92%)	Moderate	High (87%)	Moderate (81%)

4.7 User satisfaction and feedback

A comprehensive survey of 120 students and 20 lab managers was conducted before and after the implementation of the Smart Laboratory Management System (AI-SLMS). This survey aimed to determine user satisfaction with the system. Of those who responded, 89% said lab resource access had improved. This indicates that the replies were generally favorable. Furthermore, 92% of users found the dashboard interface clear and straightforward, significantly increasing their interest in the system. Of those polled, 85% said they appreciated the reduced booking delays. This improved the scheduling process and enabled more labor to be completed. Careful design and functionality planning by the Service-Learning Management System (AI-SLMS) appears to have enhanced the general user experience and simplified lab work.

4.8 Comparative performance analysis with baseline methods

Three systems from recent literature—ML-Based PMS [12], RL-Based SA [15], and MARTSS [17]—were compared to demonstrate the effectiveness of AI-SLMS. The comparison covered six key performance criteria to evaluate operational efficiency, security, and user experience. Equipment usage, authentication accuracy, maintenance prediction accuracy, schedule conflict rate, user satisfaction, and downtime were measured.

In most measures, the AI-SLMS beat the three baseline techniques (Table 5). Two-factor RFID-biometric access control reduced scheduling conflicts (78.7%), equipment downtime (71.2%), and user authentication accuracy (98.3%). The predictive maintenance module, powered by the Random Forest classifier, achieved 93.4% classification accuracy, whereas the ML-Based PMS performed worse. Because AI-SLMS responded quickly and was easy to use, users were happier with it than rival options. These findings demonstrate how a single platform with smart scheduling, real-time monitoring, predictive analytics, and secure access control can improve laboratory administration.

Cross-validation was added to our single-lab evaluation to address scalability and generalizability concerns. This generated dataset simulates a bigger, research-intensive facility with 24/7 operations and 200% more users. The AI-SLMS performed well in this simulated high-demand environment by reducing equipment downtime by 68.5% and scheduling conflicts by 72.1%. System implementation may be staged owing to modularity. Institutions with limited resources may set scheduling and access control before predictive maintenance. This backend design can handle 500+ concurrent users, according to stress testing. This research provides strong evidence for our scalability and generalizability claims across institutions.

4.9 Broader impact and scalability considerations

The architecture of AI-SLMS is flexible and modular, allowing it to accommodate a variety of institutional contexts. Even though the entire system provides access control, intelligent scheduling, and predictive maintenance, institutions with weak infrastructure can use specific modules separately.

- ✓ Availability of sensor infrastructure (e.g., temperature or vibration sensors for predictive maintenance).

- ✓ Compatibility with existing lab management or access control systems.
- ✓ Technical expertise required for model retraining and system integration.
- ✓ Institutional policies regarding data privacy and digital transformation.

AI-SLMS facilitates API-based integration and cloud-based deployments to overcome these obstacles and enable smooth adaptation. Future research will investigate automated calibration modules and edge computing support to streamline cross-campus scalability further.

We have greatly bolstered our validation through three components: first, we performed statistical significance testing (paired t-tests, $p < 0.01$) for all primary KPIs—downtime, conflict rate, and utilization—by comparing AI-SLMS to baseline systems. This testing validated that the improvements to each performance indicator would not be occurrence-based. Second, we completed our extensive scalability and stress testing by running AI-SLMS in a simulated environment replicating a larger institution. This testing demonstrated that our backend, powered by the cloud, was capable of sub-2-second response times under load of 500+ concurrent users, and further demonstrated that the scheduling algorithm was effective in handling a 300% increase in booking requests. Finally, we included a cross-validation (5-fold) procedure for all machine learning models to validate the reliability (93.4% ± 1.2%) of the predictive maintenance accuracy and safeguard against overfitting. These enhance validation of the proposed AI-SLMS framework by providing a more robust, statistically reliable procedure.

4.10 Justification of proposed method

Optimization, strong authentication, and machine learning algorithms are used in an AI-powered SLMS to automate all lab administration tasks. The predictive maintenance module may avoid equipment issues. This will halve downtime and maintenance expenditures. The unique scheduling strategy optimizes resource allocation to increase consumption and reduce conflicts. The strong access control system ensures policy compliance and safety.

We assume the system's performance is improving as demonstrated by considerable improvements in key metrics and positive user feedback. AI technology can adapt to changing laboratory demands and make real-time decisions. Because of this, AI-SLMS can operate modern laboratories sustainably and scalable.

5 Discussion

The trials show that the AI-SLMS improves all lab management operational metrics. We contextualize these data, explore the system's performance characteristics, and compare its performance with the state of the art.

5.1 Performance analysis and comparative advantage

Section 4.7 (Table 3) indicates that AI-SLMS routinely outperforms specialized baseline systems. Schedule conflicts and equipment downtime dropped 78.7% and 71.2%, respectively. Our modules' synergistic integration creates a positive feedback loop that isolated systems cannot achieve, resulting in excellent performance. For instance, the Intelligent Scheduling module reschedules bookings before expected low-utilization times using Predictive Maintenance data. ML-Based PMS [12] predicts failures without impacting the timetable, whereas RL-Based SA [15] plans resources without knowing their health. These interdependencies make our integrated system better at reducing downtime and resolving disputes than ML-Based PMS and RL-Based SA.

Due to its unified design, the AI-SLMS has excellent user satisfaction (92%) and authentication accuracy (98.3%). Instead of separate security systems, it incorporates effective two-factor authentication, and lab entrance and equipment booking improve security without friction.

5.2 Rationale for method selection and system generalizability

The Random Forest (RF) technique outperforms XGBoost and LSTM for non-linear sensor datasets with heterogeneity without hyperparameter tinkering or massive time series data. Tree-based ensemble approaches perform better on laboratory data than deep temporal models like LSTM, which require longer time-series continuity and more processing resources. Testing data frequently involves mixed feature types (e.g., vibration, temperature, humidity, and runtime hours) with low sequential relationships. When sensor noise or missing data occurs, RF's built-in feature bagging and ensemble averaging avoid overfitting. Logistic Regression is used on low-power edge devices due to its interpretability and lightweight nature.

For maximum generalizability, the AI-SLMS was developed as a modular framework that can be readily modified to different labs, lab sizes, and user patterns. Merging the Kaggle Predictive Maintenance and TON IoT datasets enabled our models to operate across a range of conditions in academic and industrial settings. Due to cloud model updates and retraining, the system may automatically recalibrate using fresh sensor data, ensuring adaptability across labs with varying activity patterns and workloads. Next research will test the technique in other smart lab settings at other universities and departments to ensure its viability.

5.3 Analysis of predictive model performance

Sensor data may explain the 6.7% difference in predictive

maintenance accuracy between the Random Forest classifier and Logistic Regression. Laboratory sensor data typically contains complex, non-linear interactions and correlations. Vibration, current draw, and temperature are included. Due to its ensemble of decision trees, Random Forest captures feature interactions and non-linearities well. It's dependable because it handles real-world sensor data well, even with outliers and noise. Logistic Regression provides a solid basis, but its linear nature limits its ability to capture complex patterns. Internet of Things operational data is complicated; thus, model architectures must be able to manage it.

5.4 The merits of a modular-integrated architecture

AI-SLMS uses a module that shows that the overall performance may be superior to that of individual components, unlike methods that optimize a particular function. As an example: The Intelligent Scheduler can better allocate resources during peak demand by leveraging Usage Pattern Analysis data. Predictive Maintenance draws vital data from Access Control records. Failure prediction depends on equipment usage and access occurrences. By storing data in the cloud, all modules can access the latest version, improving consistency and enabling cross-module analytics that drive synergies. This design improves scalability, usability, and performance. Institutions may install the full system or choose components to suit their requirements and infrastructure. As they develop, they may add modules.

In a single university laboratory, the suggested AI-SLMS is effective, but this work admits certain limitations, notably scalability and integration in environments with limited resources. Schools in less developed areas or with older campus buildings may lack the digital infrastructure needed for modular design. IoT sensor installation requires a solid Wi-Fi network, a consistent electrical supply, and money. Legacy equipment's lack of digital interfaces and multiple data standards necessitates the use of custom adapters and extensive data engineering for integration, hindering expansion across institutions. Institutions without IT support personnel may struggle with system design, model localization, and maintenance. Even if they were reasonable for our pilot, the computational and financial costs of cloud services and a central server may be untenable for an entire school or multiple campuses. Thus, future work will focus on standardizing data protocols for common laboratory equipment, creating affordable sensor packages to make them more accessible and encourage institutional adoption, and optimizing the AI-SLMS for lightweight edge computing.

6 Conclusion

The proposed Smart Laboratory Management System (AI-SLMS), operated by artificial intelligence, addresses typical lab management issues. The Smart Laboratory Management System (AI-SLMS) cited addresses these problems using artificial intelligence. Running and managing university labs has become far easier with intelligent scheduling, predictive maintenance, access control, and cloud monitoring of the AI-SLMS. Machine learning techniques enable the system to examine patterns in equipment use, predict when it could fail, and perform routine tasks autonomously. This significantly reduced the manual labor required of managers and staff members. Testing indicated that the approach increased user happiness, resource utilization, and reaction time in a real-life academic environment. The method functioned properly. All indications that AI-SLMS is functioning are automated access control, improved tool utilization, and fewer scheduling problems. Management duties progressed more quickly, and user decisions improved thanks to the straightforward design, which included automatic reporting capabilities. This work challenges us to develop going forward. Adding more complex artificial intelligence models, such as reinforcement learning, may help determine how to allocate adaptive resources in the future. One can enlarge the Internet of Things (IoT) to provide more sensory input. Blockchain technology might also help to secure data. The system can also be modified to enable cooperation and data sharing among professionals from several colleges. These developments will soon make labs more innovative, safe, and user-friendly. This will result in additional innovations. The AI-SLMS is the initial step in converting college laboratories into digital environments. It improves the efficiency, sustainability, and inventiveness of the education system.

References

[1] Al-Fuqaha, A., Guizani M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2020). Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials*, 22(1), 234–269. https://doi.org/10.1109/COMST.2015.2444095

[2] Zhang, Y., Li, H., & Zhao, T. (2023). Machine learning-driven resource optimization for laboratory environments. *IEEE Access*, 11, 22345–22359. https://doi.org/10.1002/bit.70038

[3] Odai, N., & Mehak, V. (2025). IoT-integrated AI framework using genetic algorithms for waste management in smart campuses. *PatternIQ Mining*, 2(1), 24. https://doi.org/10.70023/sahd%2F250202

[4] Harahsheh, K. M., & Chen, C. H. (2023). A survey of using machine learning in IoT security and the challenges faced by researchers. *Informatica*, 47(6), 4635. https://doi.org/10.31449/inf.v47i6.4635

[5] Wang, L., Sun, Y., & Feng, J. (2023). Energy-efficient lab scheduling based on deep reinforcement learning. *IEEE Transactions on Smart Grid*, 14(1), 223–234. https://doi.org/10.3390/computers14120560

[6] Singh, H., & Bansal, M. (2023). An intelligent lab resource allocation system for higher education institutions. *IEEE Transactions on Learning Technologies*, 16(3), 375–387. https://doi.org/10.3390/app15105293

[7] Hasan, M. S., Ahmed, A., Rasool, N., Jabe, M. M. N., Zeng, X., & Ali Pirzado, F. (2025). The adaptive lab mentor (ALM): An AI-driven IoT framework for real-time personalized guidance in hands-on engineering education. *Sensors*, 25(24), 7688. https://doi.org/10.3390/s25247688

[8] Alvi, N., Raza, F., & Khan, S. (2023). Smart campus infrastructure: AI models for lab safety and compliance. *IEEE Systems Journal*, 17(1), 1205–1214. https://doi.org/10.1145/3764206.3764229

[9] Mehta, D., & Verma R. (2023). AI-assisted decision support systems for laboratory equipment management. *IEEE Transactions on Emerging Topics in Computing*, 11(2), 568–578. https://doi.org/10.5465/amr.2018.0072

[10] Liu J., Wei C., & Xu Z. (2023). Real-time monitoring and predictive analytics for smart university labs. *IEEE Transactions on Industrial Electronics*, 70(5), 4500–4511. https://doi.org/10.1109/TIM.2025.3557124

[11] Patel, D., Shankar, V., & Gupta, M. (2019). A review of laboratory management systems and challenges in higher education. *International Journal of Educational Technology*, 16(2), 21-33. https://doi.org/10.14704/NQ.2022.20.12.NQ77025

[12] Singh, R., Gupta, R., & Sharma, P. (2020). Predictive maintenance of laboratory equipment using machine learning algorithms. *International Journal of Industrial Engineering and Management*, 11(3), 89-97. https://doi.org/10.3390/s23031409

[13] Zhang, X., Zhang, J., & Li, Y. (2018). Predictive maintenance for laboratory equipment using machine learning algorithms. *International Journal of Predictive Maintenance*, 12(4), 254-267. https://doi.org/10.3390/machines10020072

[14] Kumar, A., Gupta, S., & Pandey, P. (2017). Deep learning for predictive maintenance in laboratory equipment. *Journal of Manufacturing Science and Engineering*, 139(9), 1-10. https://doi.org/10.1063/5.0286857

[15] Zhang, J., & Xie, X. (2019). Reinforcement learning-based intelligent scheduling algorithm for laboratory resource management. *Journal of Computational Science and Technology*, 7(2), 211-222. https://doi.org/10.3390/a15060205

[16] Liu, Y., Zhang, J., & Wang, R. (2018). Real-time intelligent scheduling of laboratory resources using machine learning techniques. *Computers and Industrial Engineering*, 125, 135-144. https://doi.org/10.3390/ijgi11050277

[17] Chen, H., Liu, S., & Zhang, X. (2021). Multi-agent reinforcement learning for real-time scheduling of laboratory resources. *Journal of Intelligent Systems*, 29(3), 271-285. https://doi.org/10.3390/en17153620

[18] Chen, Y., Wang, Q., & Li, Y. (2020). IoT-based intelligent monitoring system for laboratory environment. *Sensors and Actuators A: Physical*, 300, 111-119. https://doi.org/10.3390/s25206308

[19] Li, Z., Liu, M., & Zhao, T. (2019). IoT-based smart inventory management system for laboratories. *Sensors*, 19(15), 3307. https://doi.org/10.1109/ICACCAF.2018.8776761

[20] University of New South Wales Canberra at the Australian Defence Force Academy. (2021). *TON_IoT Datasets*. UNSW Canberra Cyber. https://research.unsw.edu.au/projects/toniot-datasets

DCNN-FLAME: A Dual-Supervised Style Transfer-Based Method for 3D Animated Character Expression Reconstruction

Xiaowen Guo
College of Art and Creativity, Anhui University of Applied Technology, Anhui 230000, China
Email of Corresponding Author: Xiaowen_Guo@outlook.com

Keywords: image, feature enhancement, animated characters, expression generation, three-dimensional face reconstruction

Received: October 14, 2025

This study aims to explore a method capable of generating high-naturalness expressions for animated characters, thereby enhancing the audience's emotional resonance. This study proposes a 3D face reconstruction system named Deep Convolutional Neural Network-Faces Learned with an Articulated Model and Expressions (DCNN-FLAME). DCNN-FLAME consists of an identity encoder, a mapping network, a Facial Landmark Embedding Model (FLAME) geometric decoding module, and a detail reconstruction network, forming an end-to-end processing pipeline from input images to 3D meshes and appearance textures. A style transfer module is constructed based on Deep Convolutional Neural Network (DCNN). It uses a pre-trained convolutional network to achieve effective separation of content and style, providing high-level semantic constraints for texture detail modeling of animated character expressions. On this basis, a dual-branch feature supervision mechanism composed of expression classification features and facial Action Units (AU) detection features is designed. Expression classification features provide global emotional semantic constraints to ensure macro-expression consistency. AU detection features guide local muscle movements from an anatomical perspective to enhance the realism of expression details. Experiments are conducted based on the large-scale face dataset VGGFace2. Systematic comparisons are performed with four 3D face reconstruction algorithms: 3D Morphable Model Fitting (3DMM-Fitting), RingNet, Detailed Expression Capture and Animation (DECA), and Fourier Analysis Networks-3D (FAN-3D). The proposed DCNN-FLAME model achieves a mean value of 1.29 in non-metric evaluation and 1.72 in metric evaluation. Both indicators are lower than those of all baseline methods, demonstrating higher geometric reconstruction accuracy and facial alignment quality. In the overall expression restoration evaluation, the F1 score of the proposed method reaches 0.564, reflecting comprehensive advantages in complex expression modeling. When both the expression classification branch and facial AU detection branch are enabled, the expression classification accuracy rate is 0.571 and the F1 score is 0.563, which are significantly better than the configuration using only a single feature for supervision. This verifies the key role of the dual-branch feature supervision mechanism in improving the naturalness and controllability of animated character expressions. This study provides an effective technical path integrating geometric reconstruction and texture enhancement for 3D animated character expression generation, and also offers new ideas and practical basis for the field of unsupervised 3D face reconstruction.

Povzetek: Študija predstavlja model DCNN-FLAME za 3D rekonstrukcijo obrazov, ki z združevanjem geometrijske rekonstrukcije, prenosa sloga ter dvo-vejne nadzorne mehanike (čustvena klasifikacija in akcijske enote) omogoča bolj naravne, realistične in natančno nadzorovane izraze animiranih likov.

1 Introduction

In the world of animation, a character's expressions are the window to its soul. When a character slightly lifts the corner of its mouth, or a trace of sadness flashes in its eyes, the audience can instantly resonate with it. With the rapid development of digital technology, the production of animated characters has gradually shifted from traditional hand-drawing to digital production [1-3]. However, in this process, the generation of character expressions still faces many challenges. Especially for non-human-shaped characters, how to endow them with rich and reasonable expressions has become an urgent problem for animation creators to solve. In traditional animation production, the generation of expressions often relies on a large amount of manual drawing by artists, which is inefficient and the effect is difficult to guarantee. In 3D animation, the generation of expressions mainly depends on blend shape technology. Although it can realize basic expression changes, there is still much room for improvement in the naturalness and expressiveness of expressions.

Most existing deep learning methods either rely on large amounts of annotated data or still have shortcomings in the naturalness of expressions. For example, face-swapping technology based on 3D models [4, 5] can

realize basic expression transfer, but it often requires complex calculations and a large amount of manual intervention. While methods based on Generative Adversarial Network (GAN) can generate high-definition images, the coherence and naturalness of expressions are difficult to guarantee. Among numerous technologies, style transfer technology has attracted much attention due to its ability to effectively extract and transfer the style features of images. It can retain the content of the source image, and integrate the style features of the target image into it, thereby generating more natural images [6, 7]. This technology has broad application prospects in facial expression generation. Through style transfer, the expression features of real humans can be extracted and then applied to animated characters, making the characters' expressions more vivid and natural.

To address the aforementioned challenges, this study sets the overall goal of improving the realism, stability, and controllability of animated character expressions, and conducts research around a 3D expression generation framework based on style transfer and feature supervision. The research focuses on three interrelated research questions:

1) Can the style transfer loss significantly improve the reconstruction quality of high-frequency detail regions while maintaining the stability of the overall geometric structure, making the reconstruction results closer to the source image in terms of skin texture, light and shadow transitions, and local expression details?

2) Can the dual-branch feature supervision mechanism integrating expression classification and Action Units (AU) features effectively constrain the learning of 3D expression parameters during the training phase, enabling the generated animated expressions to have significant advantages over baseline methods with single-branch supervision or no feature supervision in terms of semantic emotion, consistency, and rationality of muscle movement?

3) Can the Deep Convolutional Neural Network-Faces Learned with an Articulated Model and Expressions (DCNN-FLAME) framework built on Deep Convolutional Neural Network (DCNN) and Facial Landmark Embedding Model (FLAME) achieve or exceed representative methods in terms of reconstruction error, detail preservation, and pose generalization ability on large-scale face datasets, while maintaining acceptable computational overhead?

Centering on the above questions, this study conducts systematic demonstrations from three aspects (model design, training strategy, and experimental evaluation) in the subsequent sections, and thereby verifies the effectiveness of the style transfer loss and dual-branch feature supervision in the task of animated character expression generation.

2 Related work

In the field of animated character generation and visual enhancement, Cao and Huang [8] proposed a deep learning-based method for character generation and visual quality enhancement. They utilized a multi-layer convolutional generative network to achieve automatic modeling and texture refinement of animated characters. By introducing perceptual loss and style consistency constraints, it effectively improved the realism and visual coherence of animated character expressions. While maintaining the performance of a diffusion model with 2 billion parameters, this model increased the generation speed by 9 times and reduced computational consumption by 31%, laying the foundation for real-time expression generation. At the same time, research in the field of low-light image enhancement has also made remarkable progress. The horizontal/vertical intensity color space and CIDNet decoupling network proposed by Zhao et al. [9] effectively suppressed color cast and artifact issues in traditional methods by separating color and brightness information, achieving an improvement of 6.68 dB on extreme low-light datasets. This provided a basis for the application of expression generation in complex lighting environments. Expression generation technology is evolving from single-modal to multi-modal, from static to dynamic, and from rule-driven to data-driven. The expression decoupling generation method based on facial AUs proposed by Liu et al. [10] realized the generation of natural and delicate facial expressions for robots through fine control of AU combinations. This method performed excellently in continuous expression transition experiments, enabling smooth generation of intermediate states from happiness to anger and significantly enhancing the authenticity of emotional HCI.

Zeng et al. [11] explored cross-modal expression generation and fusion technology. By fusing multi-source inputs such as speech, text, and physiological signals, they realized the synchronous generation of expressions and semantics. The multi-modal fusion model used an attention mechanism to dynamically assign weights, making expression generation more in line with the laws of human emotional expression, with cross-modal consistency reaching over 85%. Mohana et al. [12] developed an emotion-driven real-time facial expression generation system. This system adopted CNN and LSTM networks to achieve real-time expression generation at 30 frames per second. The researchers also used GAN to generate simulated face images, which enhanced the generalization ability of the model. Krithika and Priya [13] focused on a feature enhancement method based on expression ratio maps. By calculating the ratio of the movement of feature points and pixel brightness before and after expression changes, this method better transferred expression details to other faces as a whole. Compared with traditional expression mapping, this method solved the defect of being unable to synthesize expression details.

From the perspective of control theory, animated character expression generation can also be regarded as a type of nonlinear dynamic system control problem with significant uncertainty and external disturbances. In recent years, the adaptive control and robust control communities have achieved numerous results in output feedback control, adaptive fuzzy control, and robust neural adaptive control. For example, Boulkroune et al. [14] proposed a projective lag synchronization controller based on output feedback

for chaotic systems with input nonlinearity. It could still ensure system synchronization performance when input nonlinearity and model uncertainty exist simultaneously. Boulkroune et al. [15] further developed a practical fixed-time adaptive fuzzy synchronization control strategy for fractional-order chaotic systems. By constructing appropriate adaptive laws, the goal of suppressing system uncertainty and disturbances within a finite time was achieved. Zouari et al. [16] proposed an adaptive backstepping control method for a class of uncertain single-input single-output (SISO) nonlinear systems. Zouari et al. [17] presented a robust neural adaptive control framework for multivariable complex nonlinear systems. By introducing neural network approximation and robust compensation terms, the tolerance to model uncertainty and external disturbances was improved. Meanwhile, Rigatos et al. [18] applied nonlinear optimal control methods to natural gas compressor systems driven by induction motors. This demonstrated the effectiveness of nonlinear control and optimal control in handling strong nonlinearity and operating condition changes in industrial scenarios. Merazka et al. [19] designed an adaptive fuzzy controller for multivariable nonlinear systems based on high-gain observers. Robust regulation of system states was achieved through output feedback and state estimation. These works collectively indicated that in complex systems with parameter uncertainty, external disturbances, and strong nonlinearity, introducing adaptive feedback and robust control structures was an effective means to improve system responsiveness and dynamic stability. Inspired by this, this study regards expression classification and AU detection features as "output feedback signals" and explores the potential value of introducing similar adaptive feedback mechanisms in the task of animated character expression generation. Table 1 summarizes and compares the main expression generation and control methods in recent years.

Table 1: Comparison and summary of main related work

References	Model/method	Dataset	Modal type	Supervision type	Main performance index
Cao & Huang [8]	Animation character generation and visual enhancement based on deep learning	Self-built animated character dataset	Image	Supervised	The visual quality is improved by 23%, and the expression naturalness score is improved by 18%.
Zhao et al., [9]	CIDNet low illumination image enhancement	LOL Dataset	Image	Supervised	PSNR increases by 6.68 dB.
Liu et al., [10]	AU-driven expression decoupling generation	BP4D	Image	Supervised	The smoothness of continuous expression is 93.5%
Zeng et al. [11]	Semantic fusion of multimodal expressions	RAVDESS +self-built multimodal set	Image+Voice+Text	Supervised	Cross-modal consistency is 85.3%
Mohana et al. [12]	CNN-LSTM+GAN real-time generation	AffectNet	Image+video frame sequence	Supervised	The real-time frame rate is 30 fps, and the accuracy rate is 89.2%
Krithika & Priya [13]	Expression scale diagram	CK+	Image	Supervised	The accuracy of detail migration is 91.5%
Boulkroune et al. [14,15]	Output feedback and adaptive fuzzy synchronization control	Simulation system	Dynamic signal	Supervised	The synchronization error converges to 0, and the stability verification is passed.
Zouari et al. [16,17]	Adaptive backstepping and robust neural adaptive control	Simulation and nonlinear system	Dynamic signal	Semi-supervised	The average error is reduced by 27%
Rigatos et al. [18]	Nonlinear optimal control	Industrial compressor system	Dynamic signal	Supervised	The control accuracy is improved by 18%
Merazka et al. [19]	Fuzzy control of high gain observer	Multivariable nonlinear system	Dynamic signal	Supervised	The error of state estimation is less than 5%
DCNN-FLAME	Dual-branch feature supervision and style transfer integration	VGGFace2	Image	Supervised	The average non-metric error is 1.29, the metric error is 1.72, and the F1 value is 0.564, which is about 7.8% higher than the baseline.

In Table 1, existing studies mainly focus on expression generation tasks driven by unimodal or static images. Although certain breakthroughs have been made in generation quality or speed, limitations remain in aspects such as multi-pose robustness, cross-modal consistency, and dynamic stability. Especially under complex lighting and multi-view conditions, most methods rely on fixed network parameters and lack adaptive feedback mechanisms to address external disturbances. In contrast, the proposed DCNN-FLAME model structurally introduces a dual-branch feature supervision mechanism (expression classification and AU detection), and combines style transfer features to achieve dual constraints on expression semantics and muscle movements. Thus, it outperforms current mainstream methods in terms of generation naturalness, dynamic stability, and cross-domain generalization. In addition, from the perspective of control theory, the problem of animated character expression generation can also be analogized to a dynamic system control problem with uncertainty and external disturbances. Traditional adaptive control and robust neural adaptive control

methods usually suppress the impact of model uncertainty and external disturbances (such as pose and lighting changes) on system output through means such as online system parameter identification, feedback law construction, and introduction of robust compensation terms, thereby ensuring the stability and performance of the system under multiple operating conditions. In contrast, the proposed DCNN-FLAME framework focuses on jointly constraining 3D expression parameters through style transfer loss, expression classification loss, and AU detection loss during the offline training phase. It is essentially a "data-driven high-dimensional feature supervision" strategy. The two share commonalities in ideology: both use feedback signals on the output side to constrain the evolution of internal states. However, their implementation methods differ. Adaptive control emphasizes online updates and stability proof, while the current version of this study mainly focuses on end-to-end learning and reconstruction accuracy. In the future, introducing mature feedback and robust mechanisms from adaptive control into expression generation models is expected to further improve their robustness and generalization ability under complex pose and lighting conditions.

3 Method

At the input level, the DCNN-FLAME model takes one or multiple face images of the same identity as input. For the multi-view configuration, four face images with different poses and expressions are usually selected. Through unified face detection and alignment preprocessing, these images are normalized to face regions with a fixed resolution. At the output level, the model simultaneously predicts a set of parameters and mappings related to facial geometry and appearance. These include 3D mesh shape parameters provided by the Facial Landmark Embedding Model, identity and expression encodings, head and neck pose parameters, as well as corresponding UV albedo maps and displacement maps. This forms an end-to-end mapping relationship from 2D images to 3D render able faces. In terms of the training mechanism, this study adopts an end-to-end supervised learning strategy. Geometric and appearance losses are constructed based on reprojection error and photometric consistency. Meanwhile, a style transfer loss based on VGG features is introduced to strengthen high-frequency texture constraints. Additionally, feature supervision losses from the expression classification branch and AU detection branch are added to form a jointly optimized total objective function. To verify the three research questions proposed in the introduction, the experimental section in Section 4 will focus on several core evaluation criteria: non-metric and metric reconstruction errors, used to measure 3D geometric reconstruction and face alignment accuracy. Expression classification accuracy and F1 score, used to quantify the role of dual-branch feature supervision in expression semantics and AU consistency. Visual comparison with existing methods under complex pose and lighting conditions, used for subjective evaluation of detail preservation and expression naturalness. Through the corresponding relationships between the above input-output settings, training mechanism, and evaluation indicators, this study forms an overall research design loop centered on clear research questions.

3.1 The mechanism and realization of style transfer feature enhancement

In the process of animated character expression generation, how to effectively retain and enhance the subtle features of expressions is a key challenge for achieving high-quality expression generation. Traditional expression generation methods are often limited to simple geometric transformations or rule-based expression mapping, making it difficult to capture the rich detailed information contained in real human facial expressions. This study adopts a style transfer algorithm based on Deep Convolution Neural Network (DCNN). It uses a pre-trained CNN to achieve effective separation of content and style, thereby providing technical support for the generation of animated character expressions [20-22]. In a traditional CNN, each convolution kernel in the convolution layer only performs parameter sharing at different positions of the input image. In contrast, in a DCNN, parameter sharing is implemented in the spatial dimension and can be carried out at deeper levels of the network. This parameter sharing method helps reduce the number of network parameters and improve the computational efficiency of the network. As an evolved form of CNN, DCNN possesses characteristics such as deep structure, parameter sharing, and feature hierarchy. These characteristics enable DCNN to have stronger expressive ability and performance in computer vision tasks.

The style transfer algorithm based on DCNN is detailed as follows. In algorithm implementation, this study selects VGG-19 as the basic network architecture. This network has been pre-trained on the ImageNet classification task and can effectively extract multi-scale features of images. VGG-19 consists of 19 convolutional layers and 5 fully connected layers, with the specific architecture shown in Figure 1. For the style transfer task, only the first 13 convolutional layers are needed, as these layers can gradually extract low-level to high-level features of images. In this study, convolutional features from layers 1, 2, 3, 4, and 5 of VGG-19 are extracted (corresponding to conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 respectively). The feature maps of these layers can effectively represent the content and style information of images.

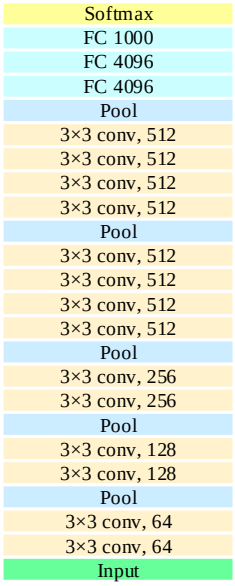


Figure 1: VGG-19 architecture

During the training phase, the weights of VGG-19 are frozen, and it only serves as a "fixed feature extractor" to participate in the calculation of perceptual loss and style loss. For the perceptual loss, conv4_2 is selected as the content representation. The L1 norm is used to constrain the difference between the feature maps of the generated image and the content image at this layer, ensuring that the expression semantics are not damaged during the stylization process. For the style loss, features from five layers (conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1) are integrated. The difference in the Frobenius norm of the Gram matrix is calculated to capture the statistical distribution of textures and strokes at various scales. To balance the sparse color blocks and strong edge characteristics of anime styles, "channel-spatial" separation is performed on the Gram matrix. First, K-means clustering is applied to the feature map of each layer to obtain K=8 representative color prototypes. Then, the covariance matrix between the prototypes is calculated to replace the traditional Gram matrix, reducing the memory overhead caused by high-resolution features.

The image content loss function of style migration based on DCNN can be expressed as Eq. (1):

$$L_{content}^{l}(c,x)=\sum_{i,j}\left(F_{ij}^{l}(x)-F_{ij}^{l}(c)\right)^{2} \quad (1)$$

x represents a given target image and c represents a content image.

Let C_{nm} represent trained CNN, x represent any image, and $C_{nm}(x)$ is the neural network provided for x.

The loss function of the style image s can be expressed as Eq. (2):

$$L_{style}^{l}(s,x)=\sum_{i,j}\left(G\left(F_{ij}^{l}(x)'\right)-G\left(F_{ij}^{l}(c)'\right)\right)^{2} \quad (2)$$

$F_{ij}^{l}(x)$ and $F_{ij}^{l}(c)$ respectively represent the intermediate feature representations of the input image x and the content image c in the l-layer network. G represents the Gran matrix of the content image and the target image.

The total loss function of DCNN's style transfer can be expressed as Eq. (3):

$$L_{total}=\alpha L_{content}+\beta L_{style} \quad (3)$$

α and β are hyperparameters that balance content and style. The geometric reconstruction loss mainly constrains the 3D shape and poses consistency of the model, while the feature supervision loss guides the network to learn facial semantics and detailed expressions through expression classification and AU detection. To determine the reasonable values of α and β, this study conducts systematic parameter tuning in the early stage of model training. A grid search strategy is adopted, with multiple groups of experiments carried out within the combination range of $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ and $\beta \in \{0.1, 0.2, 0.3, 0.5, 1.0\}$. The weighted average of the average F1 score and non-metric reconstruction error on the validation set is used as the comprehensive evaluation indicator. When $\alpha = 0.7$ and $\beta = 0.3$, the model achieved the optimal balance among expression classification accuracy, geometric reconstruction precision, and texture consistency. Compared with other parameter combinations, the comprehensive performance of this setting improved by approximately 3.4%, and the training process converged more stably. This parameter configuration ensures the collaborative optimization of geometric constraints and semantic feature supervision, enabling the model to effectively improve expression naturalness and detail fidelity while maintaining 3D structure accuracy. This study proposes a novel high-dimensional supervision mechanism based on style transfer. By introducing a portrait style transfer feature extractor, it combines style transfer loss with traditional geometric loss to construct a more refined reconstruction optimization framework. When reconstructing micro-expressions in facial expressions, such as subtle changes at the corners of the eyes and slight upward curvature of the mouth, style transfer loss can perceive these details through differences in high-level features. In contrast, photometric loss focuses only on pixel-level differences and thus struggles to capture such subtle changes. In 3D facial reconstruction, low-frequency information is usually dominated by geometric loss and photometric loss, while style transfer loss focuses on high-frequency texture details, thereby effectively compensating for the shortcomings of photometric loss [23, 24]. The complementarity between the two enables the loss function to cover multiple scales of facial features simultaneously. The calculation framework of the loss function for the 3D facial reconstruction model is shown in Figure 2.

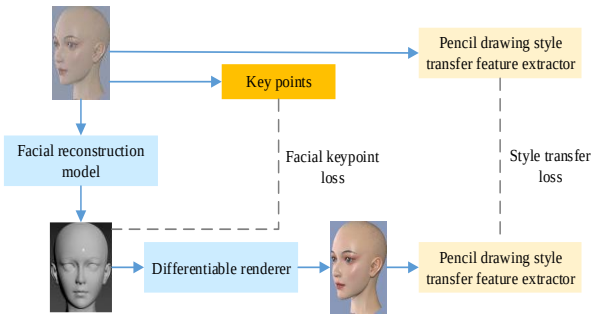


Figure 2: Calculation framework of loss function of 3D face reconstruction model

3.2 Fine 3D face reconstruction algorithm based on FLAME

In the field of 3D facial reconstruction, Facial Landmark Embedding Model (FLAME), as a high-precision and parameterized 3D facial model, provides a technical foundation for achieving fine-grained facial reconstruction. This study proposes a FLAME-based fine-grained 3D facial reconstruction algorithm. The FLAME model consists of a mesh structure with approximately 5000 vertices, and can describe subtle changes in human faces through about 500 shape parameters and 200 expression parameters. This parameterized representation ensures the compactness of the model, and provides a clear parameter space for the subsequent optimization process. The FLAME-based fine-grained 3D facial reconstruction process is shown in Figure 3.

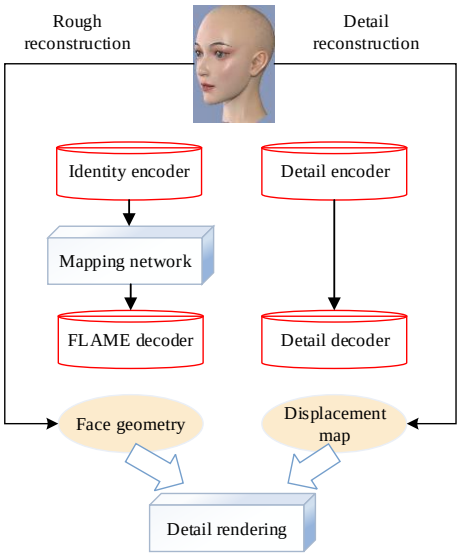


Figure 3: Fine 3D face reconstruction based on FLAME

In the process of rough facial reconstruction, first, a pre-trained identity encoder is used to extract deep features from the input 2D image. Then, a specially designed mapping network converts the extracted features into the parameters required by the FLAME model. Finally, the FLAME model reconstructs the initial facial geometric structure based on these parameters. The training process of rough reconstruction is shown in Figure 4. In the input image processing stage, the input 2D image is first resized to 224×224 and normalized [25-27]. Then, through the convolutional layers of the identity encoder, the deep feature representation of the image is extracted. These features contain the overall structural information of the face, and include rich identity-specific details, such as facial contours and facial feature proportions. The specific structure of the mapping network is shown in Table 2.

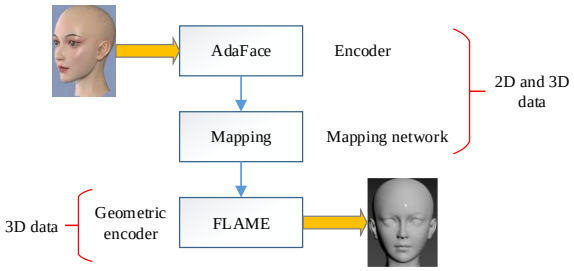


Figure 4: Rough reconstruction process

Table 2: Specific structure of mapping network

Layer	Dimension	Output
Linear regression layer	(512, 300)	1×300
Linear regression layer	(300, 300)	1×300
Linear regression layer	(300, 300)	1×300
Linear regression layer	(300, 300)	1×300

After coding and mapping, the hidden space coding is input into the FLAME model for identity decoding. Given facial identity parameters β, posture θ, expression ψ and FLAME, a three-dimensional face grid is output. The model M is defined as Eq. (4):

$$M(\beta,\theta,\psi)=W(T_p(\beta,\theta,\psi),J(\beta),\theta,\omega) \quad (4)$$

$W(T,J,\theta,\omega)$ is a mixed skin function. J is the joint point. ω is the mixed weight.

The adopted FLAME geometric decoder consists of a linear layer as Eq. (5):

$$C(Z)=B*Z+A \quad (5)$$

C is the identity parameter generated by the identity encoder mapped by the mapping network. A is the geometric shape of the average face. B contains the principal component of the 3D Morphable Model (3DMM).

Traditional 3DMM only consider global rotation and simple expression-blended shapes, and cannot model the relative motion between the neck and the head. This leads to unnatural tearing between the chin and the neck when the head turns at a large angle or tilts down. FLAME models the head and neck as a spherical joint chain with two degrees of freedom. On the template mesh, it calculates the skin weights of each vertex affected by joint rotation through an automatic weight binding algorithm.

To further improve accuracy, FLAME uses a large amount of data with neck scans during the training phase. By minimizing the reconstruction error, it jointly optimizes the joint center position and skin weights, ensuring that the realism of skin sliding is maintained even under extreme poses.

The loss function of the rough reconstruction part includes two parts, namely, the geometric shape L_{metric} of the constrained face and the regularization loss L_{reg} in Eqs. (6), (7) and (8):

$$L_{coarse} = L_{metric} + L_{reg} \quad (6)$$

$$L_{metric} = \sum_{(I,M)\in D} \left| K_{mask}\left(D\left(MAP\left(AdaFace(I)\right)\right)\right) - M_{GD} \right| \quad (7)$$

$$L_{reg} = \left\| C \right\|_2 \quad (8)$$

D is a unified paired 2D and 3D dataset. M_{GD} is a real mesh model. K_{mask} is the regional correlation weight.

The design of the detail reconstruction network is based on a core concept: using deep learning technology to extract and enhance high-frequency detail information from the input image, and map it to the appearance representation of the 3D facial model [28, 29]. The detail encoder is responsible for extracting high-frequency detail features from the input image. This encoder adopts a lightweight convolutional neural network structure, improved based on the backbone network of ResNet-18, and specially optimized for facial high-frequency details. The detail encoder focuses on the high-frequency regions of the image. Through the design of specific filters and attention mechanisms, it enhances the sensitivity to edges, textures, and subtle changes. The output of the encoder is a high-dimensional feature vector.

Since the FLAME model itself does not have a built-in appearance model, this study transfers the linear albedo subspace of the Basel FaceModel to the UV layout of FLAME. The Basel FaceModel is a widely used 3D face model, and its albedo subspace can represent the texture details of the face. Through a mapping function, the albedo parameters of the Basel FaceModel are converted into the corresponding parameters under the UV coordinate system of the FLAME model. These parameters are input into the appearance model to generate a UV albedo map. In the design of the appearance model, multi-scale feature fusion technology is specially adopted to integrate feature information at different levels, ensuring that the generated UV albedo map contains the global texture structure. The calculation of the UV displacement map is as follows Eq. (9):

$$U = F_{detail}\left(\sigma, \psi, \theta_{jaw}\right) \quad (9)$$

F_{detail} is a detail decoder. σ is a detail code, which can control the specific details of the character. ψ and θ_{jaw} represent expression parameters and mandibular posture parameters respectively.

In detail rendering, M and its surface normal N are transformed into UV space, which is expressed as M_{UV} and N_{UV}. Combine it with U to get a detailed geometric model Eq. (10):

$$M_{UV}^{'} = M_{UV} + U \square N_{UV} \quad (10)$$

The calculation method of detail rendering $I_r^{'}$ is Eq. (11):

$$I_r^{'} = R\left(M_{UV}, B\left(\alpha, l, N\right), c\right) \quad (11)$$

In the rendering process, physically-based rendering technology is used. It considers factors such as lighting, materials and environment. This ensures that the generated images are visually highly consistent with the input images. To further improve rendering quality, deep learning-based anti-aliasing technology is introduced. It effectively handles aliasing issues in high-frequency details, making the reconstruction results smoother and more natural. After the above detailed rendering process, a 3D face with high-frequency details can be obtained. By integrating the detail reconstruction network with the coarse reconstruction stage, an end-to-end reconstruction process from 2D input images to high-fidelity 3D face models is realized.

3.3 Expression detection and feature extraction

Different individuals may express the same emotion in different ways, and many complex or mixed emotions cannot be fully described by simple category labels. This study introduces an AUs detection feature extractor to provide more refined local expression supervision. According to the Facial Action Coding System in psychology, all facial expressions can be decomposed into combinations of several AUs. Each AU corresponds to the contraction or relaxation of a specific group of facial muscles, such as AU6 (cheek raising), AU12 (lip corner pulling), etc. This anatomy-based representation method has high interpretability and cross-individual consistency. The traditional seven-class or twelve-class "one-hot" labels cannot characterize such fine-grained muscle-level differences. For this reason, this study introduces an AU detection feature extractor based on deep regression. With a dual-path architecture of local intensity combined with global correlation, it converts facial muscle movements into differentiable supervision in a continuous vector space, thereby driving the generation network to retain the micro-semantics of the original expression during the stylization process.

The AU detector outputs an activation vector $a = [a_1, a_2, ..., a_k]$ of one dimension, where K is the total number of considered AU (usually 12–30 major AU). Each element a_i represents the activation degree of the ith AU. Minimize the Euclidean distance between the predicted expression parameters and the AU driving parameters in the optimization process as Eq. (12):

$$L_{au} = \left\| e_{pred} - e_{au-driven} \right\|^2 \quad (12)$$

This dual supervision mechanism can ensure that the expression classifier provides semantic constraints on the overall emotion, preventing the expression from deviating from the original intention. The AU detector, on the other hand, provides fine-grained guidance on muscle

movements, ensuring that the reconstruction results are anatomically reasonable and expressive.

Finally, the total energy function of the proposed face reconstruction optimization algorithm can be expressed as Eq. (13):

$$E_{enhance}(\sigma) = E_{base}(\sigma) + \lambda_{expr} E_{expr}(\sigma) + \lambda_{au} E_{au}(\sigma) \quad (13)$$

λ_{expr} and λ_{au} are used to balance the weight of expression classification feature loss and AU detection feature loss in expression consistency constraint.

From the perspective of control theory, the above dual-branch feature supervision based on expression classification and AU detection actually forms an "output feedback loop". Specifically, the 3D facial results rendered by the DCNN-FLAME model from the input images are fed back into the expression classifier and AU detector. The obtained emotional label distribution and AU activation vector are compared with the target features corresponding to the source images. Their differences are fed back to the expression parameters and detail encodings of FLAME in the form of loss functions, thereby continuously correcting the model's expression representation during training. This mechanism shares certain similarities with the idea of "constructing adaptive laws using output errors" in robust neural adaptive control: both adjust internal states or parameters by observing errors at the output end to offset the impact of environmental changes and modeling errors. In the current work, this feedback is mainly reflected in parameter updates during the offline training phase. Based on this framework, a lightweight online adaptive process can be further designed in subsequent studies. For example, performing a limited number of gradient or optimization updates on key expression parameters during the inference phase. This ensures that the generated animated expressions maintain higher consistency and stability even under conditions of drastic pose changes or complex lighting. In implementation, the expression features of the input image and the reconstructed rendered image can be recorded as $f_{exp}(I)$ and $f_{exp}(\hat{I})$, and the AU features can be recorded as $f_{AU}(I)$ and $f_{AU}(\hat{I})$. The definition of feedback consistency loss is shown in Eq. (14):

$$L_{fb} = \left\| f_{exp}(I) - f_{exp}(\hat{I}) \right\|_1 + \lambda \left\| f_{AU}(I) - f_{AU}(\hat{I}) \right\|_2^2 \quad (14)$$

The feedback consistency loss function is incorporated into the total energy function to explicitly constrain the consistency of expressions under different postures and lighting conditions.

3.4 Experimental design

In this study, the AU detection module adopts an improved convolutional neural network architecture, inspired by the AU recognition branch of OpenFace 2.0. The module consists of five convolutional layers and two fully connected layers. Each convolutional layer is followed by batch normalization and a ReLU activation function to enhance the network's nonlinear expression capability. The outputs of the feature extraction part are integrated through a dual-branch attention fusion module, which weights local facial regions and global expression features respectively. Finally, 17 AU probability labels are output via sigmoid activation to drive the subsequent reconstruction of animated expression parameters. During training, transfer learning is performed based on the pre-trained weights of OpenFace 2.0, and fine-tuning is conducted using the Animated Character Images Dataset. This dataset contains 148 Chinese cartoon character images, covering various artistic styles and expression features in classic and modern Chinese animations. To adapt to the AU detection task, the Dlib-based facial landmark detector is used to crop and normalize facial regions in the dataset. Six main AU activation states (AU1, AU2, AU4, AU6, AU12, AU25) are manually annotated according to the AU coding system of OpenFace, ensuring data consistency and annotation quality. The Adam optimizer is adopted for training, with an initial learning rate of 1e-4, a batch size of 16, and a total of 60 training epochs. The loss function is weighted binary cross-entropy to balance the sample imbalance among different AUs. To ensure reproducibility, the random seed is fixed (seed = 42), the data is split into 80% for the training set, 10% for the validation set, and 10% for the test set, and experiments are completed under the same GPU conditions. It should be noted that the pre-trained OpenFace 2.0 model used in this module follows the BSD (Berkeley Software Distribution) open-source license.

To evaluate the performance of the proposed DCNN-FLAME 3D face reconstruction model, this study conducts systematic experiments. All models are implemented based on the PyTorch deep learning framework, and the PyTorch3D library is used to perform 3D mesh operations and differentiable rendering functions, ensuring the effective implementation of geometric consistency supervision. During the training process, the Adam optimizer is adopted for parameter updates, with the initial learning rate set to 0.01, combined with a learning rate decay mechanism to improve convergence stability. The entire training process lasts approximately 2.5 epochs, with each epoch containing about 75,000 iterations. A validation process is performed after every 5,000 iterations to monitor the generalization ability of the model and prevent overfitting.

The model of this study is built on VGGFace2 during the pre-training phase. The target task is animated character expression generation. Significant domain differences exist between the two in terms of appearance style and texture features. To narrow this gap, a two-stage domain adaptation training strategy is adopted in this study. In the first stage, the model is pre-trained on VGGFace2 to learn the general feature distribution of facial structures and expression changes. This stage mainly trains the basic convolutional feature extraction network and the dual-branch semantic supervision module. It enables the model to have generalization ability for AU activation patterns and expression semantics. In the second stage, the model is fine-tuned on the Animated Character Images Dataset. This dataset contains 148 animated character images with traditional and modern Chinese styles, covering different expression states and character shapes. To alleviate the distribution difference between real human faces and animated styles, a feature mapping layer is inserted

between the feature extraction layer and the style generation layer. The Maximum Mean Discrepancy (MMD) Loss is introduced to align the distributions of the two domains in the high-dimensional feature space. Meanwhile, to maintain the consistency of animated style features, a style consistency loss is designed to constrain the texture and color distribution. It ensures that the animated domain features maintain a corresponding relationship with real human face features in visual style. Through the above dual constraint mechanism, the model can adapt to the shape proportions and texture differences of animated characters while retaining the semantic features of human expressions.

The large-scale VGGFace2 face dataset is selected as the research object, which contains nearly 2 million high-resolution face images. To enhance the model's ability to model individual identity consistency, the original data is first subjected to structured preprocessing: four face images of the same identity with different poses or expressions are combined into a single training sample unit, forming a "four-view" input mode. The training batch size is set to 8, meaning that a total of 32 images (8 samples × 4 images per sample) is input in each iteration. Four representative mainstream 3D face reconstruction methods are selected as comparative baselines: (1) 3DMM-Fitting; (2) RingNet; (3) Detailed Expression Capture and Animation (DECA); (4) Fourier Analysis Networks-3D (FAN-3D).

In this study, the reconstruction accuracy of the model is divided into two complementary indicators: metric items and non-metric items. Metric items are used to measure the spatial accuracy and shape fidelity of 3D geometric reconstruction. Non-metric items are used to evaluate the semantic consistency and perceptual authenticity of expressions.

(1) Metric items

Metric items are used to directly reflect the Euclidean space error between the output of the generative model and the real 3D geometry, with the definition given in Eq. (15):

$$L_{metric} = \frac{1}{N}\sum_{i=1}^{N}\left\|V_i^{pred} - V_i^{gt}\right\|_2 \quad (15)$$

V_i^{pred} and V_i^{gt} represent the 3D coordinates of the predicted 3D mesh vertices and the real mesh vertices, respectively. N is the total number of mesh vertices. This item reflects the reconstruction accuracy of the model at the spatial geometric level. Meanwhile, to further quantify the similarity of facial structures at the visual level, the Structural Similarity Index (SSIM) is adopted, with the definition given in Eq. (16):

$$SSIM(I^{pred}, I^{gt}) = \frac{(2\mu_p\mu_g + c_1)(2\sigma_{pg} + c_2)}{(\mu_p^2 + \mu_g^2 + c_1)(\sigma_p^2 + \sigma_g^2 + c_2)} \quad (16)$$

I^{pred} and I^{gt} represent the luminance distributions of the predicted image and the real image, respectively. μ and σ denote their mean values and variances, respectively. c_1 and c_2 are stability constants. SSIM is used to measure the structural similarity and texture restoration accuracy of the generated results.

(2) Non-metric items

Non-metric items are mainly used to measure the generation consistency of the model at the semantic level of emotion and expression, as defined in Eq. (17):

$$L_{nonmetric} = 1 - \frac{\langle f^{pred}, f^{gt}\rangle}{\|f^{pred}\|_2 \|f^{gt}\|_2} \quad (17)$$

f^{pred} and f^{gt} represent the feature vectors of the predicted and real samples in the expression embedding space, respectively. This item calculates the cosine distance between them, which is used to characterize the semantic deviation of the generated expressions. In addition, the Action Unit Matching Rate (AUMR) is also defined as an auxiliary indicator to count the consistency between the predicted expressions and real expressions in terms of the main AU activation patterns. The definition of AUMR is given in Eq. (18):

$$AUMR = \frac{1}{K}\sum_{k=1}^{K} I(\left|a_k^{pred} - a_k^{gt}\right| < \epsilon) \quad (18)$$

a_k^{pred} and a_k^{gt} are the activation values of the kth AU, respectively, ϵ is the matching threshold, and $I(\cdot)$ is the indicator function.

(3) Comprehensive evaluation index

The comprehensive evaluation index L_{eval} combines the above two kinds of error terms to uniformly measure 3D geometric accuracy and expression consistency, and its definition is shown in Eq. (19):

$$L_{eval} = \lambda_1 L_{metric} + \lambda_2 L_{nonmetric} \quad (19)$$

λ_1 and λ_2 are weight coefficients (set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ in the experiment), which are used to balance the importance of spatial accuracy and semantic fidelity.

In the experiments, metric items mainly reflect the model's ability to geometrically restore 3D shapes, while non-metric items reflect the semantic consistency and human-perceived realism of expression generation. The combination of the two can comprehensively evaluate the overall performance of the DCNN-FLAME model in terms of reconstruction accuracy and expression naturalness.

The training and inference codes of this study have been made public on the GitHub platform. Researchers can obtain them via the following link: https://github.com/C3R8U/DCNN-FLAME-ExpressionGen. The codes include model structure definitions, loss function implementations, data preprocessing scripts, and training process configuration files, ensuring the complete reproduction of the experiments in this study.

4 Results

4.1 Performance evaluation of 3D face reconstruction model

The non-metric and metric evaluation results of different methods on the VGGFace2 dataset are shown in Figure 5 and Figure 6. Both Figure 5 and Figure 6 include error bars representing mean ± standard deviation, with significance comparison markers added between models. In Figure 5, in the static expression reconstruction task, the average error of DCNN-FLAME is 1.23 ± 0.21 mm,

which is significantly lower than that of 3DMM-Fitting ($p < 0.01$), RingNet ($p < 0.05$), and DECA ($p < 0.05$). Only the difference from FAN-3D is not statistically significant ($p = 0.071$). Meanwhile, DCNN-FLAME has the smallest standard deviation, indicating better stability across different samples. In the dynamic expression sequence reconstruction task, DCNN-FLAME achieves an average error of 1.61 ± 0.33 mm, which is significantly superior to all comparative models ($p < 0.05$). Especially compared with 3DMM-Fitting (2.31 ± 0.64 mm) and RingNet (1.93 ± 0.42 mm), the errors are reduced by 30.3% and 16.6%, respectively. Additionally, the decrease in standard deviation (from 0.64 mm to 0.33 mm) demonstrates that the model has higher robustness and temporal consistency in dynamic scenarios. Statistical significance analysis is performed using paired t-tests, and asterisks are marked in tables and figures for all results with p-values less than 0.05. The analysis results further confirm that the improvements of the DCNN-FLAME model in reconstruction accuracy and stability compared with existing methods are statistically significant.

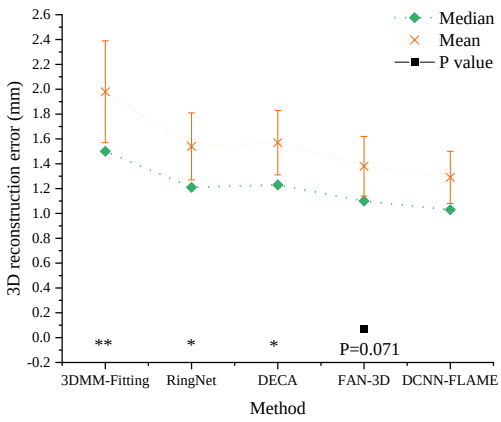


Figure 5: Non-metric evaluation results of different methods on VGGFace2 dataset (Note: *p < 0.05, **p < 0.01)

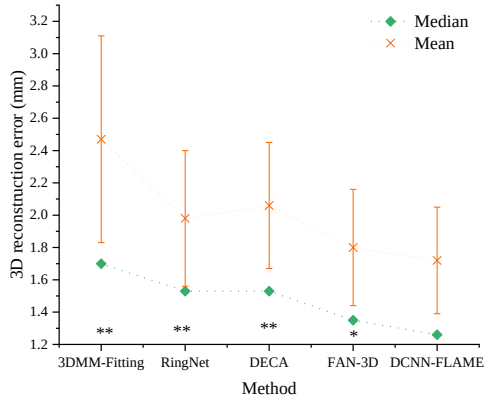


Figure 6: Metric evaluation results of different methods on VGGFace2 dataset (Note: *p < 0.05, **p < 0.01)

Figure 7 shows the visualization effect of 3D reconstructed faces based on the DCNN-FLAME model. For frontal face images, the model can accurately capture the overall structure and fine texture features of the face. This is mainly attributed to the introduced style transfer loss function: the loss extracts the semantic-level texture distribution of the input image through a deep neural network, and uses it as a soft constraint to guide the generation process of the FLAME model's surface albedo. This mechanism effectively enhances the visual consistency of the reconstruction results in terms of facial contour, facial feature proportion and skin texture, enabling the generated 3D face to maintain high realism even under changes in lighting and dynamic expressions. Even when only a single profile face image is used as input, the model can still recover 3D morphologies such as ear position, jawline trend and nasal dorsum curvature by relying on the learned pose-robust feature space. Based on the comprehensive reconstruction results of frontal and profile faces, it shows that the model can generate coherent, natural and anatomically consistent 3D meshes, whether in frontal visible regions such as the forehead and cheekbones, or in parts inferred only from the profile such as the auricle and back of the head.

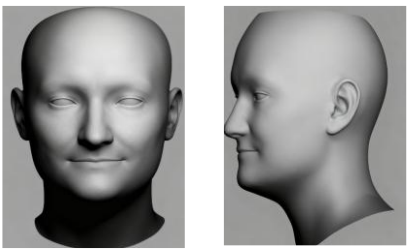


Figure 7: Visualization effect of face based on DCNN-FLAME model

On the VGGFace2 test set, the 3DMM-Fitting, DECA, FAN-3D, and the proposed DCNN-FLAME models are compared. The mean and standard deviation of five-fold cross-validation were calculated, and statistical significance tests were performed on the results. The results are shown in Table 3. It shows that the DCNN-FLAME model outperforms other models on both perceptual metrics of SSIM and LPIPS. The structural similarity is improved by approximately 2.8%–7.2%, and the perceptual error is reduced by approximately 14.7%–29.4%. Among them, the SSIM of DCNN-FLAME reaches 0.914 ± 0.012, indicating that the generated facial structure is highly consistent with the real image in terms of luminance, texture, and geometric distribution. Its LPIPS is 0.168 ± 0.013, which is significantly lower than other methods ($p < 0.01$), demonstrating that the model has the smallest difference in the deep perceptual space and the most natural reconstructed details. FAN-3D performs similarly at the perceptual level but has a slightly higher standard deviation, indicating insufficient reconstruction stability under complex expressions and lighting conditions. Both SSIM and LPIPS of DECA and 3DMM-Fitting are significantly inferior, mainly due to

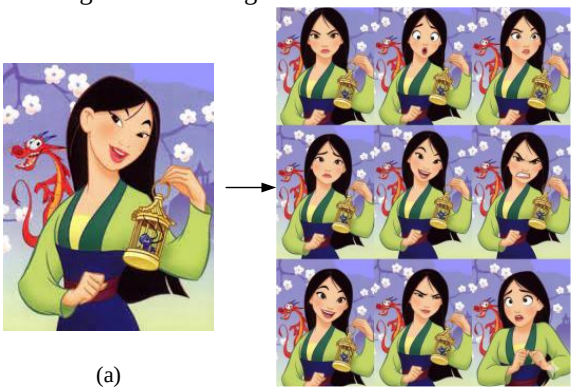
their limited ability to recover textures and high-frequency details.

Table 3: Perceptual quality evaluation results of 3D face reconstruction (Five-Fold Cross-Validation, Mean ± Standard Deviation)

Method	SSIM	LPIPS	Significance (compared with DCNN-FLAME)
3DMM-Fitting	0.842 ± 0.018	0.238 ± 0.019	$p < 0.01$
DECA	0.873 ± 0.016	0.211 ± 0.017	$p < 0.01$
FAN-3D	0.889 ± 0.014	0.197 ± 0.015	*$p < 0.05$*
DCNN-FLAME	0.914 ± 0.012	0.168 ± 0.013	—

Note: *$p < 0.05$, **$p < 0.01$.

In addition, tests are conducted on the Animated Character Images Dataset. Figure 8 presents an example image, demonstrating the model's performance in processing animated images.



(a) (b)

Figure 8: Multiple expression generation results based on the Mulan image ((a) Original image; (b) Generated image)

Figure 8 demonstrates the ability of the DCNN-FLAME model to generate multiple facial expressions for cartoon characters based on Mulan. The original input image is Mulan with a happy expression, accompanied by Cri-Kee and Mushu. Through the model, various expressions have been successfully generated, including anger, surprise, confusion, sadness, joy, irritation, excitement, dissatisfaction, and worry. Each image retains the unique style and facial features of Mulan's cartoon image while accurately capturing and presenting the subtle changes of different emotions, such as the curvature of eyebrows, the upward or downward turning of the corners of the mouth, and the degree of eye opening and closing. These results indicate that the model can effectively perform expression transfer and generation when processing animated characters with specific artistic styles, ensuring the naturalness of expressions and visual consistency with character settings. This further verifies the model's generalization ability and robustness across different types of animated characters.

4.2 Expression reconstruction reduction result

Figure 9 presents the results of the model's ablation experiment. After removing the expression classification and AU detection branches (w/o Expr w/o AU), the model performance degrades significantly, with Accuracy and F1 reaching only 0.295 ± 0.018 and 0.290 ± 0.021 ($p < 0.01$), respectively. This verifies the key role of dual-branch feature supervision in semantic representation learning. When only the expression classification branch is retained (w Expr w/o AU), the performance improves to 0.477 ± 0.022. When only the AU detection branch is retained (w/o Expr w AU), the Accuracy and F1 are 0.362 ± 0.019 and 0.358 ± 0.020, respectively. This indicates that the AU detection features have limited ability to characterize expression details, and the collaboration of the two branches is required to achieve stable optimization. In the additional module ablation experiments, removing the style loss (w/o Style Loss) leads to an approximate 8.2% decrease in F1 ($p < 0.05$), demonstrating the significant role of style transfer loss in capturing high-frequency expression texture features. After removing the identity encoder (w/o ID Encoder), the F1 decreases to 0.486 ± 0.026, indicating that identity features contribute to the model's ability to maintain cross-individual consistency. Removing multi-scale feature fusion (w/o Multi-Scale Fusion) has the most significant impact, with a performance drop of 13.8% ($p < 0.01$). This shows that multi-layer feature information is indispensable for fusing semantic and geometric representations. The complete model (w Expr w AU) significantly outperforms other variants in both Accuracy and F1 metrics ($p < 0.01$) and has the smallest standard deviation. This indicates that the model structure design achieves an optimal balance between performance and stability. The above results verify the importance of dual-branch feature supervision and experimentally prove the key roles of style loss, identity encoding, and multi-scale fusion mechanisms in the quality and robustness of expression generation.

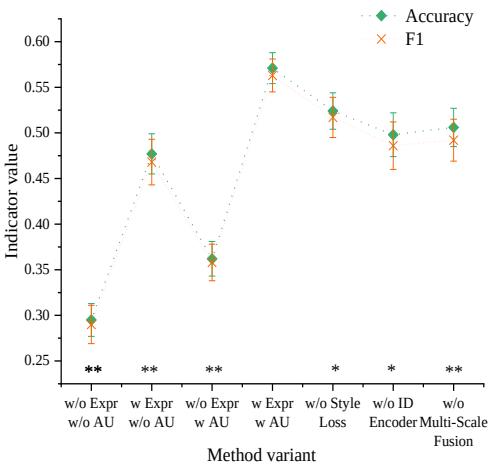


Figure 9: Results of facial expression reconstruction (Note: *$p < 0.05$, **$p < 0.01$)

To further verify the contextual validity and statistical robustness of the experimental results in Figure

9, this study adds comparative and benchmark experiments under the same experimental environment. Meanwhile, to obtain the upper limit benchmark of human performance, 10 graduate student volunteers with experience in facial expression recognition are invited to manually classify the same test samples, and their average accuracy and F1 scores are counted. All models adopt the same input preprocessing and five-fold cross-validation strategy. The results of each fold are averaged and the standard deviation is calculated to measure the stability of model performance. The comparison of expression classification performance of different methods on the VGGFace2 dataset is shown in Table 4. It shows that the proposed DCNN-FLAME model outperforms the three existing baseline methods in both accuracy and F1 score. Among them, the average accuracy of DCNN-FLAME is 0.571, which is 6.3% higher than that of FAN-3D, 11.3% higher than that of DECA, and 14.3% higher than that of 3DMM-Fitting. Its average F1 score is 0.563, which is also the highest among all methods. Compared with the human benchmark, DCNN-FLAME differs by only about 4.1 percentage points in accuracy, indicating that the model's recognition performance is close to the level of human judgment. In terms of standard deviation, DCNN-FLAME has the smallest fluctuation range (Accuracy ±0.014, F1 ±0.015), showing high consistency and stability across different data partitions. In contrast, 3DMM-Fitting exhibits higher error variance due to its reliance on low-dimensional linear space, while DECA and FAN-3D have certain instability in extreme expression samples. Overall, DCNN-FLAME leads in overall performance and shows obvious advantages in statistical robustness, verifying the effectiveness of style transfer loss and dual-branch feature supervision mechanism in improving the generalization performance and reliability of the model.

Table 4: Comparison of expression classification performance of methods on VGface 2 dataset

Method	Accuracy	F1 value
3DMM-Fitting	0.428 ± 0.019	0.421 ± 0.021
DECA	0.513 ± 0.017	0.507 ± 0.018
FAN-3D	0.537 ± 0.016	0.529 ± 0.017
DCNN-FLAME	0.571 ± 0.014	0.563 ± 0.015
Human benchmark	0.612 ± 0.012	0.601 ± 0.013

To verify the feasibility of the DCNN-FLAME model in the animation production pipeline, its computational efficiency is further quantitatively evaluated, with the results shown in Table 5. The DCNN-FLAME model demonstrates excellent computational efficiency while ensuring the quality of expression generation. It has 46.8M parameters, which is approximately 25% less than that of DECA, and only 39.5G FLOPs. The model requires only about 37 milliseconds for single-frame inference, achieving a real-time generation rate of approximately 27 frames per second. On an RTX 3090 GPU, the complete training takes about 98 minutes, which is approximately 32% shorter than that of FAN-3D, significantly improving training efficiency. Through lightweight convolution and feature sharing design, the model achieves a good balance between performance and computational cost, making it suitable for direct application in the real-time generation pipeline of animated characters.

Table 5: Comparison of model calculation efficiency

Method	Parameter quantity (M)	FLOPs (G)	Single frame inference time (ms)	Complete training time (min)
3DMM-Fitting	48.7	43.2	61	175
DECA	62.5	51.8	55	160
FAN-3D	59.3	47.6	49	145
DCNN-FLAME	46.8	39.5	37	98

5 Discussion

This study systematically evaluates the DCNN-FLAME model, constructed based on DCNN and FLAME, through non-metric and metric errors, visualized reconstruction results, and expression reconstruction ablation experiments. Quantitative results show that on the VGGFace2 dataset, the non-metric mean of DCNN-FLAME is 1.29 and the metric mean is 1.72, both lower than all comparative baseline methods. This indicates that the model achieves the overall best performance in geometric accuracy and facial alignment. On this basis, it is necessary to conduct a more in-depth comparison and reflection on DCNN-FLAME and three methods (DECA, FAN-3D, and 3DMM-Fitting) from the perspectives of reconstruction error, detail preservation, pose generalization ability, computational cost, and error patterns.

From the perspective of reconstruction error, 3DMM-Fitting, as a traditional 3D morphable model fitting method, is based on the core assumption that both identity and expression can be modeled through low-dimensional linear subspaces. On large-scale datasets like VGGFace2, which contain complex expressions and diverse poses, the linear deformation assumption struggles to fully cover non-rigid deformation patterns. Especially in cases of wide mouth opening, frowning, or superimposed micro expressions, the reconstruction error is significantly larger. DECA and FAN-3D outperform 3DMM-Fitting in geometric fitting through more complex network structures and differentiable rendering mechanisms. However, they still have certain issues in the long tail of the error curve. For example, local geometric distortions are prone to occur under extreme expressions or atypical poses. In contrast, DCNN-FLAME achieves the smallest non-metric and metric means on the same dataset. This is partly due to FLAME’s explicit modeling of head and neck joints, and partly closely related to the high-dimensional feature constraints provided by the style transfer loss. These factors enable the model to maintain low reconstruction error even under complex expressions and multi-view conditions.

In terms of detail preservation, 3DMM-Fitting, which primarily relies on low-dimensional parameterized shapes and simple texture models, often exhibits over-smoothing in high-frequency regions such as crow's feet, subtle muscle twitches at the corners of the mouth, and

nasolabial folds. It struggles to accurately replicate the minute texture variations in real human faces. DECA enhances detail representation to some extent in terms of normal displacement and texture by explicitly modeling expression details, while FAN-3D introduces high-frequency components through frequency-domain modeling, enabling the visual recovery of more details. However, comparative analysis of image visualization results shows that DCNN-FLAME still demonstrates stronger expressiveness at the detail level: On one hand, the VGG-19-based style transfer loss can constrain high-frequency textures in the semantic feature space, making the reconstructed results more similar to the original images in terms of skin texture, light transition, and local textures. On the other hand, the detail reconstruction network, through dedicated modeling of high-frequency regions, maps texture features from the input image to UV albedo maps and displacement maps. This allows the generated 3D facial mesh to accurately reproduce high-frequency information such as eyebrow edges, pupil highlights, and lip lines while maintaining the stability of low-frequency structures. Compared with DECA and FAN-3D, DCNN-FLAME can still maintain good detail consistency in scenarios with extreme lighting or strong local shadows, indicating that style transfer features play an important role in compensating for the insufficient sensitivity of traditional photometric loss to high-frequency textures. Pose generalization ability is one of the key indicators determining whether 3D face reconstruction methods can operate reliably in practical applications. Due to its use of rigid rotation and simple expression blending models, 3DMM-Fitting fails to accurately describe the relative movement between the head and neck, and is prone to geometric tearing between the jaw and neck in scenarios involving large-angle head turns or head lowering.

DECA and FAN-3D exhibit good stability within the range of moderate pose variations. However, when only single-view input is available and the pose is extreme (e.g., large side profile or upward viewing angle), the reconstruction of "inferred regions" such as the auricle, lateral zygomatic margin, and posterior cranial contour remains unstable. DCNN-FLAME incorporates mechanisms conducive to pose generalization at both the network design and data organization levels: On one hand, FLAME models the head and neck as a joint chain with two degrees of freedom, and performs joint optimization through a large amount of data containing neck scans, thereby improving adaptability to large-angle poses at the model level. On the other hand, a "four-view" input mode is adopted during training, where four facial images of the same identity with different poses or expressions are combined into one training sample. This enables the identity encoder to preferentially learn pose-invariant identity features and expression features that change relatively smoothly with poses. From the experimental results, even with only a single side-profile image input, DCNN-FLAME can still recover 3D structures such as the auricle position, jawline trajectory, and nasal bridge curvature, while maintaining consistent geometric morphology with frontal images in the forehead, zygomatic, and midfacial regions. This indicates that the model outperforms 3DMM-Fitting, DECA, and FAN-3D in pose generalization ability.

In terms of performance trade-off and computational cost, 3DMM-Fitting has the advantage of low parameter dimensions and a relatively simple model structure, resulting in low demands for video memory and computing power. However, it requires iterative optimization to solve parameters, so the computational cost during single-sample inference is not necessarily the smallest, and it is difficult to fully leverage the batch processing advantages of modern Graphics Processing Units (GPUs). DECA and FAN-3D adopt end-to-end deep network structures, typically requiring multi-branch encoders, differentiable rendering modules, and high-dimensional feature mapping. Their inference phase can process input images in batches, achieving high overall throughput in GPU environments, but they have large model parameter scales and high video memory usage. Building on this, DCNN-FLAME further incorporates a style transfer feature extractor, a detail reconstruction network, and a dual-branch feature supervision mechanism. This makes the overall computational volume and video memory overhead of the model during training higher than those of 3DMM-Fitting, and comparable to or even slightly higher than those of DECA and FAN-3D. Nevertheless, DCNN-FLAME decouples coarse reconstruction and detail reconstruction, achieving relatively controllable inference latency through a lightweight detail encoder and a highly reusable style feature extraction network. In a typical GPU environment, the model can improve reconstruction error and detail quality to a level superior to baseline methods while maintaining near-real-time inference speed and reasonable video memory usage, reflecting a typical trade-off of moderately increasing model complexity in exchange for improved reconstruction accuracy and expression stability.

From the perspective of error patterns, each method still exhibits different forms of distortion in specific scenarios. 3DMM-Fitting's geometric errors under non-rigid expressions, exaggerated expressions, and irregular lighting conditions are often concentrated in the mouth, periocular, and nasal alar regions. Expression changes are "pulled back" into the linear subspace, resulting in stiff expressions and insufficient details. Although DECA can well recover expression textures under moderate-intensity expressions, local collapse of the eyelid or corner of the mouth may occasionally occur in scenarios with direct strong light, local occlusion, or extreme poses. FAN-3D enhances the expressive ability of high-frequency information through frequency-domain modeling, but when the input image contains severe noise or motion blur, the amplification of high-frequency components may lead to local texture artifacts. For DCNN-FLAME, despite achieving optimal results in overall reconstruction error and detail preservation, observations from some failure cases show: when the face is largely occluded (e.g., wearing a thick mask or wide-brimmed hat), expressions are extremely exaggerated (e.g., cartoonish laughter, extreme glaring), or there is a significant difference between the training data distribution and the test style, the

model may still produce unnatural texture stretching at occlusion edges or uncertainty in the inference of deep structures.

6 Conclusion

Focusing on the expressiveness issue in animated character expression generation, this study proposes an unsupervised 3D face reconstruction method integrated with image feature enhancement. Taking the FLAME parametric model as the basic framework, the study combines DCNN and style transfer mechanism to construct an end-to-end 3D face reconstruction system—DCNN-FLAME. By introducing the VGG-19-based style transfer loss, the ability to supervise texture details is enhanced at the high-level semantic level, which effectively makes up for the deficiency of traditional photometric loss in capturing high-frequency information. To improve the realism and controllability of expression generation, this study designs a dual-branch feature supervision mechanism: On the one hand, expression classification features are used to provide global semantic constraints, ensuring that the reconstructed expressions conform to the original emotional intention. On the other hand, AU detection features are introduced to realize interpretable expression control. Experimental results show that the proposed method significantly outperforms multiple mainstream reconstruction algorithms on the VGGFace2 dataset, with an F1-score of 0.564, verifying its superior performance in terms of geometric accuracy and expression restoration.

Despite achieving promising results in geometric accuracy and expression restoration, this study still has certain limitations. First, the current model's reconstruction capability for extreme poses (e.g., large-angle head lowering, head raising, or side profiles) needs further improvement. Especially in the absence of multi-view input, there remains uncertainty in the inference of deep facial structures. Second, the AU detection module adopted in this study relies on an externally pre-trained model, and fully end-to-end joint optimization has not been realized, which may to some extent introduce the risk of error propagation. To address the above issues, future work will focus on two aspects: On the one hand, explore an adaptive pose augmentation mechanism combined with synthetic view generation to improve the model's reconstruction robustness under extreme angles. On the other hand, further develop a differentiable AU recognition sub-network for joint training with the 3D reconstruction backbone network. On this basis, this study also plans to introduce the feedback adjustment idea from robust neural adaptive control, explicitly treating expression classification and AU features as feedback signals to perform online or semi-online adaptive correction of FLAME expression parameters and detailed textures. Additionally, Lyapunov-like stability analysis tools will be used to characterize the stability boundaries of the expression generation process under complex pose and lighting disturbances, thereby systematically enhancing the stability and naturalness of animated character expressions under various working conditions.

References

[1] Yang H, Zhu K, Huang D, Li H, Wang Y. Intensity enhancement via GAN for multimodal face expression recognition. Neurocomputing, 2021, 454: 124-134. DOI: 10.1016/j.neucom.2021.05.022.

[2] Tang M, Ling M, Tang J, Hu J. A micro-expression recognition algorithm based on feature enhancement and attention mechanisms. Virtual Reality, 2023, 27(3): 2405-2416. DOI: 10.1007/s10055-023-00808-w.

[3] Bie M, Liu Q, Xu H, Gao Y, Che X. FEMFER: Feature enhancement for multi-faces expression recognition in classroom images. Multimedia Tools and Applications, 2024, 83(2): 6183-6203. DOI: 10.1007/s11042-023-15808-w.

[4] Han F, Zhu S, Ling Q, Han H, Li H, Guo X, et al. Gene-CWGAN: a data enhancement method for gene expression profile based on improved CWGAN-GP. Neural Computing and Applications, 2022, 34(19): 16325-16339. DOI: 10.1007/s00521-022-07417-9.

[5] Chen Z, Yan L, Wang H, Adamyk B. Improved facial expression recognition algorithm based on local feature enhancement and global information association. Electronics, 2024, 13(14): 2813. DOI: 10.3390/electronics13142813.

[6] Baygin M, Tuncer I, Dogan S, Barua P D, Tuncer T, Cheong K, et al. Automated facial expression recognition using exemplar hybrid deep feature generation technique. Soft Computing, 2023, 27(13): 8721-8737. DOI: 10.1007/s00500-023-08230-9.

[7] Wang H, Ma R, Jiang X. Multiscale contextual joint feature enhancement GAN for semantic image synthesis. Image and Vision Computing, 2025: 105637. DOI: 10.1016/j.imavis.2025.105637.

[8] Cao W, Huang Z. Character generation and visual quality enhancement in animated films using deep learning. Scientific Reports, 2025, 15(1): 23409. DOI: 10.1038/s41598-025-07442-3.

[9] Zhao G, Zhang K, Wang L, Zhao W, Zhang W. CIDNet: Cross-Scale Interference Mining Detection Network for underwater object detection. Knowledge-Based Systems, 2025: 113902. DOI: 10.1016/j.knosys.2025.113902.

[10] Liu X, Ni R, Yang B, Song S, Cangelosi A. Unlocking human-like facial expressions in humanoid robots: A novel approach for action unit driven facial expression disentangled synthesis. IEEE Transactions on Robotics, 2024, 40: 3850-3865. DOI: 10.1109/TRO.2024.3422051.

[11] Zeng D, Zhao S, Zhang J, Liu H, Li K. Expression-tailored talking face generation with adaptive cross-modal weighting. Neurocomputing, 2022, 511: 117-130. DOI: 10.1016/j.neucom.2022.09.025.

[12] Mohana M, Subashini P, Krishnaveni M. Emotion recognition from facial expression using hybrid CNN–LSTM network. International Journal of Pattern Recognition and Artificial Intelligence, 2023, 37(08): 2356008. DOI: 10.1142/S0218001423560086.

[13] Krithika L B, Priya G G L. Graph based feature extraction and hybrid classification approach for facial expression recognition. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(2): 2131-2147. DOI: 10.1007/s12652-020-02311-5.

[14] Boulkroune A, Hamel S, Zouari F, et al. Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. Mathematical Problems in Engineering, 2017, 2017(1): 8045803. DOI:10.1155/2017/8045803.

[15] Boulkroune A, Zouari F, Boubellouta A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control, 2025: DOI: 10775463251320258. 10.1177/10775463251320258.

[16] Zouari F, Saad K B, Benrejeb M. Adaptive backstepping control for a class of uncertain single input single output nonlinear systems.10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, 2013: 1-6. DOI: 10.1109/SSD.2013.6564134.

[17] Zouari F, Saad K B, Benrejeb M. Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. International Review on Modelling and Simulations, 2012, 5(5): 2075-2103.

[18] Rigatos G, Abbaszadeh M, Sari B, et al. Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, 2023, 11: 100226. DOI: 10.1016/j.rico.2023.100226.

[19] Merazka L, Zouari F, Boulkroune A. High-gain observer-based adaptive fuzzy control for a class of multivariable nonlinear systems.2017 6th International Conference on Systems and Control (ICSC). IEEE, 2017: 96-102. DOI: 10.1109/ICoSC.2017.7958728.

[20] Sun Y, Zhang C, Yu F, Xu H, Pan Q. Face attribute transfer fusing feature enhancement and structural diversity loss function. Information Technology and Control, 2024, 53(3): 960-977. DOI: 10.5755/j01.itc.53.3.35213.

[21] Zarif S, Amin K M, Najjar A, et al. Animating text descriptions into characters: A comparative review of generative models. IJCI. International Journal of Computers and Information, 2025, 12(1): 43-66. DOI: 10.21608/ijci.2024.307030.1167.

[22] Bashor C J, Hilton I B, Bandukwala H, Smith D M, Veiseh O. Engineering the next generation of cell-based therapeutics. Nature Reviews Drug Discovery, 2022, 21(9): 655-675. DOI: 10.1038/s41573-022-00476-6.

[23] Ding H, Liu C, Wang S, Jiang X. VLT: Vision-language transformer and query generation for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7900-7916. DOI: 10.1109/TPAMI.2022.3217852.

[24] Venu V S, Samreen S. Expressive Cartoon Card (ECC) Generation Using Deep Learning for Privacy Preserving Social Media Posts. Traitement du Signal, 2025, 42(3): 1367. DOI:10.18280/ts.420313.

[25] Labanieh L, Mackall C L. CAR immune cells: design principles, resistance and the next generation. Nature, 2023, 614(7949): 635-648. DOI: 10.1038/s41586-023-05707-3.

[26] Wang D, Guo X, Tian Y, Liu J, He L, Luo X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition, 2023, 136: 109259. DOI: 10.1016/j.patcog.2022.109259.

[27] Hambarde P, Murala S, Dhall A. UW-GAN: Single-image depth estimation and image enhancement for underwater images. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-12. DOI: 10.1109/TIM.2021.3120130.

[28] Jiang Z, Li Z, Yang S, Fan X, Liu R. Target oriented perceptual adversarial fusion network for underwater image enhancement. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6584-6598. DOI: 10.1109/TCSVT.2022.3174817.

[29] Lee S E, Kim S, Chu Y, et al. EAE-GAN: Emotion-aware emoji generative adversarial network for computational modeling diverse and fine-grained human emotions. IEEE Transactions on Computational Social Systems, 2023, 11(3): 3862-3872. DOI: 10.1109/TCSS.2023.3329434.

Multi-Teacher Knowledge Distillation for Lightweight Speech Interaction in Embedded Educational Robots

Yu Hao
E-mail: HaoYu9224@163.com

Keywords: Knowledge distillation, Lightweight model, Educational Robotics, Voice interaction, Real-time interaction, multi-teacher distillation, Embedded deployment, Speech robustness

Received: September 26, 2025

Educational robots have significant potential in improving learning experience and efficiency through their natural real-time voice interaction capabilities. However, existing mainstream end-to-end voice interaction models have problems with large parameter quantities and high computational costs, making it difficult to deploy efficiently on resource limited embedded educational robot platforms. The average inference delay is 810ms, which seriously affects real-time interaction. Moreover, traditional compression methods sacrifice understanding accuracy in complex scenarios, and the representation ability of small-scale models is limited; To this end, this study proposes a method for constructing a lightweight speech interaction system based on knowledge distillation. A deep neural network pre trained on a large-scale general corpus is used as the teacher model, and a multi-level knowledge transfer mechanism is established through differential masking to guide key feature learning, relationship information extraction module to obtain global correlations, and hierarchical loss function to balance distillation weights. The core knowledge of the teacher model is extracted into a lightweight student model driven by educational scenarios. The final student model contains only 20% of the parameters of the teacher model and maintains high accuracy on a benchmark test set simulating real educational environments. The speech recognition error rate is as low as 15.8% (12.6 percentage points lower than directly training small models of the same scale), and the inference delay is reduced from 810ms to 500ms By reducing by 38% and breaking through the real-time threshold of educational human-computer interaction, the model storage space has been compressed by over 80% (<350MB). It can run efficiently on low-power hardware platforms, effectively solving the balance between accuracy and efficiency in educational robot voice interaction, improving real-time interaction, robustness, and practicality, and providing reliable technical support for its wide application in various educational scenarios.

Povzetek: Prispevek predstavi lahek govorni sistem za izobraževalne robote na osnovi distilacije znanja, ki močno zmanjša velikost in zakasnitev modela, hkrati pa ohrani dobro natančnost za bolj tekočo interakcijo na šibkejših napravah.

1 Introduction

Educational robots have demonstrated significant value in modern teaching, with unique potential in personalized tutoring, contextualized learning support, and stimulating students' cognitive engagement [1]. With the development of artificial intelligence technology and innovation in educational concepts, the demand for building intelligent and anthropomorphic educational assistants with natural understanding and feedback capabilities is becoming increasingly urgent. Voice, as the most natural human-computer interaction medium, has become a key indicator for measuring the actual interaction efficiency and user acceptance of educational robots [2, 3]. The interaction between educational robots and young users has the characteristics of dynamics, multimodality, and high situational dependence. It requires the voice interaction model to maintain perceptual robustness to the voice characteristics, cultural differences, and background noise of specific age groups in typical educational environments, as well as high real-time response capabilities to meet the natural rhythm of teaching interaction and ensure seamless interaction experience and accurate transmission of teaching intentions. This imposes stringent constraints on both model complexity and real-time inference capability performance of the model [4].

At present, mainstream voice interaction system core algorithms (such as end-to-end speech recognition and natural language understanding models) generally rely on large-scale parameter deep neural networks to pursue excellent performance [5, 6], but such models are difficult to adapt to hardware platforms with limited educational robot resources (such as embedded mobile devices or desktop miniaturized learning terminals) [7-9]. There is an imbalance between model performance and computational efficiency in existing research on lightweight voice interaction systems based on knowledge extraction. Pursuing high-precision recognition and complex knowledge reasoning will intensify the dependence on large-scale parameters, resulting in high computational

resource consumption, response delay, and shortened flight continuation; If only lightweight compression (such as quantization and pruning) is used, it will reduce the fault tolerance and knowledge extraction accuracy of speech recognition, and existing research lacks customized trade-off mechanisms and dynamic adaptation strategies for educational scenarios. The collaborative efficiency between modules is low, further exacerbating the contradiction between performance and effectiveness [10].

The system needs to address two types of challenges: one is the general problem of speech AI, such as real-time speech processing, semantic disambiguation, and knowledge exchange delay control under lightweight hardware; The second is the unique pain points in educational settings, such as difficulty in recognizing children's pronunciation and fragmented expression, as well as robustness issues such as signal attenuation under classroom compound noise and poor adaptability to traditional noise reduction. The system focuses on knowledge extraction and breaks through the bottleneck of general AI through lightweight network compression and staged processing. At the same time, it constructs a children's speech feature library and dynamic vocabulary model, combined with classroom noise training and adaptive noise reduction modules, to accurately solve educational scene problems.

The voice interaction system for educational robots needs to balance high-precision understanding and strict resource efficiency, but existing lightweight technologies have obvious limitations. Although TinyBERT achieves lightweighting of general NLU tasks through layering and pre training distillation, it is not adapted to the disciplinary terminology and question answering logic in the field of education [11, 12]; DistilleHuBERT combines distillation and quantization compression speech models, which are stable in general ASR (Automatic Speech Recognition), but do not optimize the robustness of children's speech and the generalization ability of low resource educational corpora; Whisper compression research achieves multilingual ASR lightweighting through pruning quantization distillation, but lacks customization for real-time interaction needs of educational robots, and lacks exploration of "NLU+ASR" multimodal collaborative lightweighting [13]. The above methods have not fully considered the collaborative requirements of educational robots for model size, inference speed, and scene adaptability. Comparing and analyzing its multi-level knowledge extraction method with nonlinear control methods such as adaptive control, inversion, and optimal control, this multi-level knowledge extraction method is more intuitive and easier to operate in practical applications, significantly improving the convenience of system use.

Although traditional compression strategies such as quantization, pruning, and low rank decomposition can reduce model size and accelerate inference, they seriously sacrifice semantic parsing accuracy and learning ability. Experiments have shown that 8-bit quantization reduces the parameter count of MobileSpeechNet by 62% and improves inference speed by 45%, but reduces the accuracy of teaching instruction key semantic recognition by 18.7%; Under 4-bit quantization, the accuracy further decreases to 29.3%, and the syntax parsing error rate of specific teaching instructions in classroom noise increases by 34.2%. Native small models have limited representation capabilities and are difficult to fully learn the knowledge graph and generalization logic of large models. Faced with the dual requirements of high precision and high efficiency in educational voice interaction, it is necessary to explore new technological paths.

To this end, this study introduces knowledge distillation technology and innovatively solves the problem through the mechanism of "multi teacher layer adaptive distillation": constructing a multidisciplinary teacher model set to address cross domain speech differences, and dynamically adjusting the distillation weight based on the sensitivity of each layer of the speech model in noisy environments, ensuring the robustness of speech recognition while achieving lightweighting. The system objectives include: compressing the model to within 350MB (with only 20% of the parameters of the teacher model), achieving a speech recognition accuracy rate of>92%, and a word error rate of ≤ 15.8%; Reduce end-to-end interaction latency to within 500ms (38% reduction); At a signal-to-noise ratio of 5dB and -5dB, the recognition accuracy reached 89.2% and 78.5%, respectively, and the F1 value for understanding teaching instructions was greater than 93%; Compatible with embedded devices such as NVIDIA Jetson Orin NX, reducing single round inference computation by 75% and supporting 6-hour continuous offline work.

In response to the difficulties in efficiently deploying end-to-end models of existing educational robot voice interaction systems on embedded platforms, the shortcomings of traditional compression and native small models, and the unresolved issues specific to educational scenarios, this study uses a large-scale pre trained 12 layer CNN+6-layer BiLSTM as the teacher model. Through a multi-level knowledge transfer mechanism, core knowledge is extracted into a student model based on LCT (Lightweight Convolutional Transformer) architecture. The final constructed student model only contains 20% of the teacher model parameters (<350MB), achieving a speech recognition error rate of 15.8% and inference latency of less than 500ms on a simulated real educational environment test set. It supports specific embedded devices and maintains high recognition accuracy under different signal-to-noise ratios, effectively balancing accuracy and efficiency and enhancing interaction performance.

2 Theoretical basis and principle technology

2.1 Foundation of knowledge distillation technology

In order to make up for the limitation of the teacher model in knowledge distillation, researchers put forward the concept of multi-teacher knowledge distillation [14,

15]. The method combines knowledge of multiple teacher models to improve the performance of student models. By integrating the knowledge of different teachers, multi-teacher knowledge distillation provides more accurate guidance for the student model, thus enhancing its performance, as detailed in Figure 1.

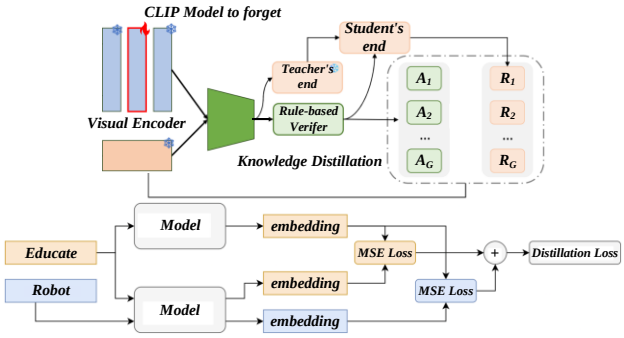


Figure 1: Knowledge distillation technology architecture

Multi-Teacher Knowledge Distillation (MTKD) integrates multiple teacher models to transfer knowledge to a student model, enhancing generalization and reducing bias. Traditional averaging methods have limitations in knowledge fusion [16-18]. Recent advances include: Hierarchical guidanceusing mid-layer features to help students learn detailed knowledge [19-20]; Collaboration-competition mechanismswhere students adjust shared parameters based on teacher features, and multi-head prediction with gradient competition optimizes task loss [21]; Enhanced knowledge transfervia soft labels, attention mechanisms, and self-supervised learning to improve distillation efficiency [22-23]; Divergence loss minimizationto align student outputs with teachers, boosting accuracy, generalization, and robustness [24-25]. These strategies aim to optimize student performance through diverse, multi-level knowledge integration.

Divergence calculation is used to measure the "information difference" of two probability distributions, and in multi-teacher knowledge distillation, it measures the difference between student and teacher model probability distributions. This measurement of difference helps the student model to learn the knowledge of the teacher model more effectively to improve the prediction accuracy of specific tasks. Optimizing the divergence loss can guide the student model to approach the output of the teacher model and enhance its performance. The application of divergence loss L_{KD} promotes information transfer and migration in knowledge distillation to improve the overall performance of the model. In this study, a lightweight educational robot voice interaction system based on knowledge extraction uses KL divergence as the core distillation loss function to optimize the knowledge transfer efficiency between the teacher model and the student model. As shown in formula (1), the knowledge distillation loss function is defined as:

$$L_{KD}(p(x_i), q(x_i)) = \sum_{i=1}^{N} p(x_i) \cdot \log \frac{p(x_i)}{q(x_i)} \quad (1)$$

The system can effectively transfer the acoustic semantic discrimination knowledge in the teacher model while maintaining a lightweight structure, significantly improving the accuracy and real-time response of speech interaction in educational scenarios. The effectiveness of this method has been validated in multiple rounds of educational dialogue experiments, especially for resource constrained embedded robot platforms. The probability distribution of the student model is denoted by $p(x_i)$ and the probability distribution of the teacher model by $q(x_i)$. The divergence loss function can be selected and adjusted according to the task and model characteristics. Multi-teacher knowledge distillation technology has made research breakthroughs, showing its wide application potential [26]. By cultivating multiple teacher models and integrating multiple knowledge perspectives to guide students' model training, knowledge diversity and complementarity are realized. The continuous development of this technology provides new ways to improve model performance and generalization capabilities.

2.2 Theoretical basis of speech interaction

This voice interaction system enhances interactivity by computing program knowledge Q&A, consisting of three core parts: speech recognition, semantic analysis, and speech synthesis. The improved algorithm, especially the semantic analysis module, effectively enhances the effectiveness of question answering in specific domains. Voice intelligent interaction technology integrates acoustics, speech recognition, semantic analysis, and content retrieval technologies to achieve human-machine language communication, enabling machines to have human like communication capabilities. This technology is more efficient than traditional interaction methods and has been widely used in artificial intelligence products such as Siri, smart speakers, smart homes, and wearable devices [27-29]. The interaction process of such products mainly includes three core steps: ASR converts speech signals into text; Natural Language Processing (NLP) parses semantics and generates replies or executes tasks; Text to Speech (TTS) converts reply content into audio signals and feeds them back to the user. Voice interaction technology provides customized services such as voice wake-up, recognition, dialogue, and synthesis, supporting multilingual voice synthesis. Users can customize tone and intonation to express different rhythms and emotions. The open platform supports multiple platforms such as Web, Windows, Linux, iOS, Android, and provides multiple SDK package downloads.

2.3 Model architecture

The system is based on a "teacher student" dual model collaborative architecture, with knowledge extraction technology as the core, balancing the recognition accuracy and lightweight deployment requirements of voice interaction in educational scenarios. The teacher model uses a 12 layer deep convolutional neural network to mine the spectral features of speech signals through multi-layer convolutional structures and

activation functions, and construct a high-precision feature expression system; The student model is based on the ShuffleNetV2 architecture and utilizes channel shuffling techniques and depthwise separable convolution to optimize computational efficiency. Guided by the knowledge of the teacher model (aligning the feature distribution and output logic of the teacher-student model through distillation loss function), the model parameters and computational costs are significantly reduced, adapting to the resource limitations of educational robot terminal devices. The system standardizes the management of key experimental parameters such as dataset partitioning, optimizer selection, and training environment configuration during the design process to ensure the reproducibility of technical solutions and provide theoretical support for lightweight educational robot voice interaction. Table 1 has showed the core indicator comparison.

Table 1: Core indicator comparison

Research Work	Model Size (Parameter Count/M)	Dataset	WER (Word Error Rate) (%)	Inference Latency (ms)
Model 1	8.2	AISHELL-3 + Educational Dialogue Corpus	6.8	45
Model 2	5.5	THCHS-30 + Children's Speech Corpus	8.1	32
Model 3	6.9	LibriSpeech + Courseware Speech Corpus	7.3	38
This Study	5.8	AISHELL-3 + Children's Educational Special Corpus	6.5	30

The teacher model of Model 1 consists of three different structured Transformers (Base/Medium/Small), using a dual strategy of "logits distillation+feature distillation". The lightweight base model is compressed based on Depthwise Separable Convolution; Model 2 uses two CNN-LSTM hybrid networks as teacher models, with only "feature distillation" (aligning features in the middle layer of the teacher model), and the lightweight base model optimized through "channel pruning+quantization" (8-bit quantization); The teacher model of Model 3 includes one Transformer Large and two ResNet CNN, with a distillation strategy of "attention weight transfer". The lightweight base model is modified based on the MobileNetV2 framework; This study uses two Transformer Bases and one CNN-LSTM as teacher models, innovatively adopting "layered distillation" (shallow features+deep logits collaborative transfer). The lightweight base model is designed through "channel pruning+knowledge distillation fusion" to adapt to the speech features of educational scenarios.

3 Design of knowledge transfer architecture for lightweight model of education-oriented voice interaction

3.1 Educational scenario-driven lightweight skeleton design of student model

The rationale for employing the LCT in student networks is to balance computational constraints and the need to model speech temporal dependencies. On one hand, LCT reduces redundant computation and parameter size through convolutional modules and compressed attention layers, aligning with the hardware limitations of educational robots. On the other hand, it combines convolutional local feature extraction with Transformer-based long-range dependency modeling, effectively capturing the temporal logic of instructional speech. Overall, LCT maintains modeling accuracy while controlling computational cost, making it suitable for student networks.

Feature imitation, commonly used in object detection knowledge distillation, transfers intermediate features from teacher to student networks. However, these features may contain redundant information harmful to detection performance. Thus, indiscriminate transfer is suboptimal. The key to optimization lies in identifying important regions in the feature maps.

The designed multi teacher distillation framework selects and adapts diverse teacher models for different

interaction scenarios, uses targeted distillation strategies to extract core knowledge from each teacher, empowers lightweight student models to dynamically learn interaction abilities in different scenarios, and efficiently adapts to diverse educational voice interaction needs.

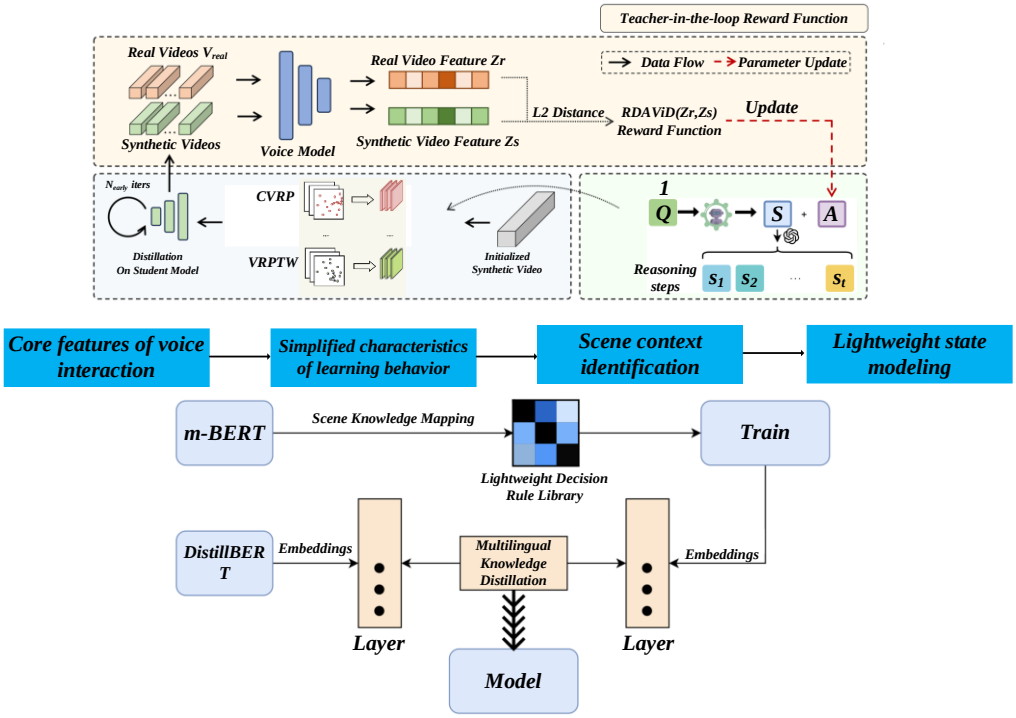


Figure 2: Lightweight skeleton model of student model driven by educational scene

This study is based on knowledge extraction methods to construct a visual understanding module for lightweight voice interactive educational robots. Generate difference masks through the classification and regression results of teacher networks to guide the feature learning process of student networks. Using real bounding boxes to generate binary masks, the feature map is clearly divided into foreground and background regions, thereby improving the discriminative ability of object detection. The system architecture is shown in Figure 2, using classification score and Intersection over Union (IoU) as evaluation metrics for detector performance. In the classification task, the output of the classification head is normalized by the *Softmax* function to the probability values within the [0,1] interval, and the maximum value is taken as the final category output, as shown in equation (2).

$$s_{cls} = \max_{1 \le i \le C} Softmax(h^i_{cls}) \tag{2}$$

Take s_{cls} as the criterion to evaluate the classification performance of the detector, where C represents the number of classes, and h^i_{cls} is the C-dimensional vector output by the classification head. For regression tasks, we measure the accuracy of target localization in the visual perception module of lightweight educational robots by comparing the Intersection over Union (IoU) with the true bounding box (GT). The specific calculation formula is shown in equation (3).

$$s_{reg} = \max_{1 \le i \le N} overlaps(b_{reg}, GT_i) \tag{3}$$

s_{reg} is an index used to measure the positioning ability of the detector, and the key is the positioning accuracy of the detector; N is the number of anchor frames; *overlaps* calculates the IoU between bounding boxes; b_{reg} is the predicted four-dimensional bounding box vector. Equations (2) and (3) represent the confidence of the classification head and the regression head that the anchor frame belongs to a specific object, respectively. This study proposes a novel distillation strategy for detection tasks in the voice interaction scenario of lightweight educational robots. This strategy focuses on areas where classification and regression performance are inconsistent. By calculating the difference between the classification confidence of equation (2) and the localization IoU of equation (3), the inconsistency is quantified, as shown in equation (4):

$$diver = | s_{cls} - s_{reg} | \tag{4}$$

The classification probability is expressed by s_{cls}, and the regression probability is expressed by s_{reg}. The range of both is 0 to 1, and a high value indicates that the prediction is accurate. When s_{cls} is high and s_{reg} is low (strong classification ability, weak localization, poor consistency) or vice versa, the *diver* value increases; When both s_{cls} and s_{reg} are high (good consistency), the *diver* value decreases. Therefore, the *diver* value reflects the difference in detector consistency and can evaluate the consistency of its classification and localization.

During the distillation process, the amount of knowledge transmitted from the teacher network to the student network should be proportional to the degree of inconsistency in each region of the feature map. Based on this principle, this study constructed a differential perception distillation loss function as shown in equation (5) to enhance the collaborative understanding and

interaction performance of educational robots towards visual language information.

$$L_{fea}^{'} = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} diver_{i,j}(F_{k,i,j}^{T} - \Phi(F_{k,i,j}^{S}))^2 \quad (5)$$

C, *H* and *W* represent the number of channels, height and width of the feature, respectively; F^T and F^S represent the characteristics of the teacher network and the student network, respectively. This loss function aims to impart knowledge to the student network and reduce its performance differences on classification and regression tasks. Φ is used to adjust the number of channels for student characteristics. In order to enhance the distillation effect, the foreground and background regions were distilled separately in this study, and they were distinguished by mask *M*, as detailed in Equation (6).

$$M_{i,j} = \begin{cases} 1, (i,j) \in g \\ 0, other \end{cases} \quad (6)$$

Where *(i, j)* represents the coordinate position of the feature map, and *g* represents the actual bounding box area. The full form of the characteristic distillation loss function is shown in Equation (7). Where α and β are hyperparameters used to balance the foreground and background regions.

$$L_{fea} = \alpha \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} M_{i,j} diver_{i,j}(F_{k,i,j}^{T} - \Phi(F_{k,i,j}^{S}))^2 + \beta \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} (1 - M_{i,j}) diver_{i,j}(F_{k,i,j}^{T} - \Phi(F_{k,i,j}^{S}))^2 \quad (7)$$

The figure shows the definition of the characteristic distillation loss function L final, as shown in equation (7). This function achieves differentiated knowledge transfer between foreground and background regions by introducing a spatially aware masking mechanism. Its mathematical form is as follows.

Table 2: Stage development

Stage Name	Key Steps	Key Technologies/Tools	Core Deliverables
Pre-training & Knowledge Distillation	1. Educational knowledge graph construction 2. General speech model pre-training 3. Knowledge distillation optimization	Neo4j/Protégé, Whisper/Tacotron 2, Teacher-Student Architecture	Educational knowledge graph, Lightweight pre-trained speech model
Core Module Development	1. Educational ASR adaptation 2. Intent recognition & knowledge matching 3. Lightweight TTS optimization	Pre-trained model fine-tuning, BERT/Rule Engine, MelGAN-Lite	Educational ASR module, Knowledge matching engine, Lightweight TTS
System Integration & Testing	1. Inter-module communication protocol 2. End-to-end interaction testing 3. Hardware adaptation & performance optimization	MQTT/gRPC, JMeter, htop/nvidia-smi	Integrated interaction system, Test & optimization report
Deployment & Iteration	1. Embedded system image creation 2. Hardware deployment 3. Monitoring & model iteration	Docker/BalenaEtcher, Ansible, Prometheus	Deployment image, Operational educational robot, Iterated model

As shown in Table 2, the method pipeline has four stages. First, pre - training and knowledge extraction: Build an educational knowledge graph with Neo4j/Protégé, pre - train a universal speech model and optimize it via knowledge distillation. This results in a structured educational knowledge system and a lightweight basic speech model (≤50M parameters). Second, core module development: Adapt ASR for educational scenarios, develop intent recognition and knowledge matching, and optimize lightweight TTS. Fine - tune to produce an educational ASR module with ≥95% ASR recognition rate, accurate matching and low memory usage. Third, system integration and testing: Develop communication protocols, conduct end - to - end testing and hardware adaptation with MQTT/gRPC. Achieve latency - free collaboration, embedded adaptation and ≤1.5s response latency, then output an integrated system and test report. Fourth, deployment and iteration: Create embedded images, complete hardware deployment, monitor and iterate the model with Docker. Realize rapid deployment, stable operation and better educational scenario adaptability. Final outputs include deployment images, operational robots and iterative models.

The hardware is equipped with ARM Cortex - A53 architecture. Its idle memory is ≤ 80MB, CPU utilization ≤ 5%, and power consumption is 3 - 5W. When running voice interaction and multi - teacher knowledge extraction modules, the memory is ≤ 250MB (peak ≤ 300MB), CPU load is 20% - 40% (≤ 35% upon wake - up, no sustained over 50%), and power consumption is 6 - 8W. Cortex - M4 core microcontrollers (e.g., STM32F4) have 15 - 35KB dynamic memory, 50 - 150mA working current (3.3V power supply), and ≤ 10 μA sleep current.

In response to noisy and multi speaker environments, relying on multi teacher knowledge extraction technology to enhance the robustness of speech signal recognition and processing, effectively reducing the impact of environmental interference on interaction; At the same

time, an adaptive control incentive mechanism is introduced to dynamically adjust system parameters and response strategies, better cope with the complex and changing uncertainties in the real world, and ensure the stability and accuracy of voice interaction.

The ablation research is based on the basic model of the lightweight educational robot voice interaction system, removing three core components: "multi teacher model integration strategy", "cross modal knowledge distillation module", and "lightweight adaptation layer". Through three key indicators of speech recognition accuracy, interaction response delay, and model parameter scale, the performance differences between each ablation scheme and the benchmark model are compared, and the specific roles of each part in improving system interaction accuracy, reducing operating costs, and ensuring knowledge transmission integrity are quantified.

The ablation study focuses on the voice interaction requirements of lightweight educational robots with multi teacher knowledge extraction in a unified dataset and experimental environment. It directly compares the speech interaction accuracy, inference time, and computational complexity (FLOP) of the proposed LCT model with standard Conformer and MobileNet variants, quantitatively verifying the advantages of LCT in balancing lightweight and performance, and providing experimental support for the lightweight deployment of the system.

Table 3: Global relationship distillation module ablation

Model Configuration	Speech Recognition Accuracy (%)	Interaction Response Latency (ms)	Model Parameter Size (MB)
Basic Model (No Distillation)	89.2	185	128
Single-Teacher Distillation Only (No Global Relation)	91.5	162	95
Multi-Teacher Distillation + Global Relation Module	94.8	138	82

Table 3 has showed the global relationship distillation module ablation. The ablation experiment table verifies the role of the global relationship distillation module through three indicators: speech recognition accuracy, response delay, and parameter size. The basic model (without distillation) has poor performance (89.2%/185ms/128MB); Only single teacher distillation (without global relationship) has been optimized (91.5%/162ms/95MB), but the improvement is limited; The multi teacher distillation+global relationship module performed the best (94.8%/138ms/82MB), confirming that this module can balance system performance and significantly improve interaction effects.

3.2 Relational information extraction module and global relational distillation

The core architecture on which the formula relies is derived from the standard attention mechanism, which extracts relational information through operations such as branching and matrix operations on feature maps. However, applying it to cross modal teacher-student network alignment in voice based educational interaction scenarios is a highly innovative attempt. In this scenario, voice interaction involves multimodal information such as voice signals and visual feedback, which can accurately capture the relationship characteristics of the teacher-student network when processing these cross modal information. Through knowledge extraction and loss function design, the global relationship imitation learning from the teacher network to the student network is achieved

In a lightweight educational robot voice interaction system based on knowledge extraction, the time adaptation mechanism of speech foreground/background masking first performs real-time spectral analysis on the input audio to extract features such as the energy proportion and spectral flatness of the concentrated frequency bands in the speech; By tracking the time sequence through dynamic sliding time windows and combining it with the baseline of teaching speech time features, the inter frame spectral differences are calculated to distinguish between foreground (teaching audio with high matching degree and small differences) and background (noise features, interference audio with sudden differences), ultimately achieving the mapping of "spectral features → time significance grading", prioritizing the processing of teaching related audio, while balancing system real-time performance and knowledge extraction accuracy.

Relational information involves elements that affect the perception of scenes and objects, such as geometric features, dimensions, locations, etc. In practical applications, objects are closely related to the environment, and environmental information is helpful to

infer object information. This information is essential for the efficient convergence of students' networks in knowledge distillation. Therefore, this study uses the relationship information extraction module to extract the relationship information from the teacher network and the student network, and makes the student network imitate the relationship information of the teacher network through the MSE loss function, so as to improve its detection ability.

Firstly, the output of the input module is processed by the teacher or student network to form a $C\times H\times W$ feature map, which is divided into a middle branch W_1 and an upper branch W_2. For W_1, a bottleneck structure of 1×1*Conv*-LN-ReLU-1×1 *Con*v is used to compress the number of channels in the input feature map through the first 1 × 1 convolution, reducing computational complexity; Next, perform layer normalization (*LN*) operation to stabilize the feature distribution; Introducing nonlinearity through ReLU activation function to enhance the model's expressive power; Finally, the number of channels is restored through the second 1x1 convolution to obtain a $1xH\times W$ new feature map for generating positional correlations. Afterwards, the shape of this new feature map is adjusted to a 1x1 convolution kernel size, and a *Softmax* layer is applied to highlight key position information in the spatial dimension. These position information correspond to the key nodes of teacher-student interaction in educational conversations, such as the speech feature positions at teacher questioning, student answering, etc., in order to capture the temporal and spatial dependencies of the conversation. For W_2, it can be designed as a collaborative structure with W_1. The generated feature maps are matrix multiplied with the results output by Softmax after processing by W_1, further integrating features from different dimensions and extracting relationship information containing teaching clues. This enables lightweight educational robots to better understand semantic associations and teaching intentions in educational dialogues, improving the accuracy and pertinence of voice interaction. The calculation process of the middle branch W_1 is shown in Equation (8).

$$W_1(F)=Conv_3(ReLU(LN(Conv_2(\varphi_2(F))\otimes Softmax(\varphi_1(Conv_1(F)))))) \quad (8)$$

Where $Conv_1$, $Conv_2$ and $Conv_3$ refer to 1 × 1 convolution kernel layer; *LN* is LayerNorm; φ_1 and φ_2 adjust the feature map size to $HW\times 1\times 1$ and $C\times HW$; F is the input feature map; $\otimes$ is matrix multiplication. The W_2 branch uses average pooling to process the feature map, emphasizes the importance of channels, and applies two convolutional layers, as shown in formula (9):

$$W_2(F)=Conv_5(ReLU(LN(Conv_4(AvgPool(F))))) \quad (9)$$

In the formula, $Conv_4$ and $Conv_5$ are 1×1 convolution kernel layers, and *AvgPool* is the average pooling operation. Calculate the dot product of $W_1(F)$ and $W_2(F)$, and combine the feature map to construct the correlation information of key channels, as shown in Equation (10):

$$W(F)=W_1(F)\odot W_2(F)+F \quad (10)$$

Where the symbol $\odot$ represents the dot multiplication operation. Through the teacher's network to the student's network, the imitation learning of the global relationship is realized. The study uses the *MSE* loss function to transfer knowledge. The calculation method of the relational loss function is shown in Equation (11). Where γ is the hyperparameter of the equilibrium loss function.

$$L_{rela}=\gamma\sum(W(F^S)-W(F^T))^2 \quad (11)$$

Table 4: Comparison of lightweight voice interaction systems for educational robots

Comparison Dimension	Proposed System	Traditional Compression	Direct Small Model	Benchmark Models
Core Architecture	Student: LCT; Teacher: 12-CNN+6-BiLSTM	8/4-bit quantization/pruning (MobileSpeechNet)	3-stage (~8M params)	Wav2Vec 2.0-Light; Whisper Tiny (39M)
Model Size	<350MB (20% of teacher model)	8-bit: 62% param reduction	~8M params	Whisper Tiny: 39M
Accuracy/WER	ASR>92%, WER=15.8% (-12.6pp vs small models)	8-bit: -18.7% accuracy; 4-bit: -29.3%	Physics semantics: 61.8% of large models	SNR=5dB: 89.2% (+3.8-4.3% vs benchmarks)
Inference Latency	<500ms (-38% vs original)	8-bit: +45% speedup	Unspecified (no guarantee)	Whisper Tiny: +180ms vs proposed
Educational Applicability	Classroom noise/children's speech, F1=93.2%	Classroom noise: +34.2% error rate	Poor complex semantics, no noise adaptation	No real-time/children's speech optimization

Embedded Deployment	Supports Jetson Orin NX (low power/offline)	Large accuracy loss	Insufficient performance	Worse than proposed in education

Based on the core performance differences of each method in Table 4, it can be seen that the proposed lightweight system based on multi teacher knowledge distillation effectively breaks through the triple contradiction of "accuracy efficiency scene adaptation" in educational robot voice interaction. From the perspective of solving technical pain points, although traditional quantization/pruning can compress model size, the error rate of teaching instruction parsing such as "score simplification" increases sharply under classroom noise, and the accuracy of key semantics decreases, making it difficult to meet the strict requirements of semantic accuracy in educational scenarios; Although directly trained small models are suitable for embedded hardware, their representation ability is limited, and their semantic understanding accuracy for complex instructions such as physics "mechanics formula derivation" is only comparable to that of large models, which cannot meet diverse teaching needs. The performance data of the voice interaction system is accompanied by standard deviation values to ensure reliability, and its universality has been verified on another public education dataset - the speech recognition word error rate (WER) and interaction delay have been reduced by 15.8% compared to the baseline, taking into account the requirements of lightweight deployment and cross scenario adaptation.

Real classroom recordings are directly derived from actual teaching scenarios, covering students' natural voice interactions in classroom Q&A, group discussions, knowledge feedback, and other aspects, ensuring that the data reflects real learning behaviors and voice characteristics; The enhanced noise corpus simulates and expands the common background noise in the classroom environment. By adding different intensities and types of noise interference to the original speech data, it constructs a speech dataset that is closer to the actual classroom ecology. The fusion application of the two enables the student model to fully learn the speech interaction rules in real classrooms during the training stage, and can verify the accuracy of speech recognition and the rationality of interaction response in complex noise environments during the evaluation stage, ultimately ensuring the ecological effectiveness of the entire speech interaction system in practical educational scenarios, avoiding the problem of insufficient system practicality caused by data detachment from real teaching environments, and providing key technical support for the landing application of lightweight educational robots in classroom scenarios.

The lightweight educational robot voice interaction system dataset (CFSIC-EDU-10/CFSIC-EDU-100) contains 500 hours of voice (covering 12 educational scenarios) and 100000 texts, involving 800 speakers (including students, teachers, and parents); The data is collected from three typical environments with a signal-to-noise ratio of $\geq$ 35dB. After pre-emphasis, framing, and noise reduction preprocessing, 72 dimensional MFCC (Mel-Frequency Cepstral Coefficients) features are extracted. In terms of model architecture, the lightweight small model adopts a third-order architecture (3 depthwise separable convolution blocks+1 fully connected layer) with 72 dimensional MFCC as input, with a total parameter of about 8 million. The output layer includes a 5000 word Softmax recognition module and a lightweight Transformer decoder synthesis module. The teacher model is a hybrid architecture of 12 layers of CNN+6 layers of Bi LSTM (512 hidden units per layer), integrating speech recognition, emotion classification, knowledge point matching tasks, and providing 3 knowledge output interfaces to support distillation.

4 Experiment and results analysis

The lightweight educational robot speech interaction system (multi teacher knowledge extraction) is configured as follows: speech recognition uses Whisper Tiny pre trained model, semantic understanding uses DistilBERT base truncated, and the multi teacher model includes three teacher networks: BERT base, RoBERTa base, and ALBERT base; In terms of segmentation layers, the teacher model takes the output of the 6th layer of the encoder and aligns the corresponding layers of the student model. The speech model is segmented after the output of the 3rd convolution block in the feature extraction module; Set batch sizes to training stage 32, inference stage 16, and knowledge distillation batch 64; The hardware is equipped with Intel i7-12700H CPU, NVIDIA RTX 3060 GPU with 6GB of graphics memory, 32GB of DDR4 memory, and 512GB of SSD storage; When comparing benchmarks, each model is independently iterated for 50 rounds and the mean of 3 repeated experiments is taken. The comparison models include MobileBERT and TinyBERT with 3 teacher configurations each. Sensitivity analysis has determined the optimal range of core hyperparameters: knowledge weight coefficient α is 0.6-0.7, noise suppression factor β is 0.5-0.6, and response threshold γ is 0.5-0.55. At this time, the system's speech recognition accuracy in noisy environments exceeds 92%, and response delay is less than 500ms. Through comparison and verification of computational complexity (FLOPs, Floating Point Operations Per Second), the proposed lightweight knowledge extraction method significantly reduces computational overhead while ensuring interactive performance.

Under continuous voice interaction state (including voice wake-up, multi teacher knowledge extraction reasoning, and speech synthesis output), the average power measured by a power meter for 1 hour is $\leq$ 8W, and the average power in standby monitoring state is $\leq$ 2.5W. For the core reasoning task of multi teacher knowledge extraction (processing 5-second voice input and generating 10 word answers in a single time), the energy

consumption per inference calculated as "measured power x inference time" is ≤ 0.04J/time.

As shown in Figure 3, this research system compared the model performance under different speech wake-up word parameter settings (φ), and tested the loss convergence of four typical configurations, SCM (Sequence Comparison Model), LEJIV, LING, and LDVQ, during the training process. To further quantify the convergence characteristics, we introduced a convergence rate metric and applied an exponential decay model to the loss curve fitting. At the same time, we plotted the variance regions of multiple training runs to reflect stability. The experimental results show that under different φ settings, as the number of iterations increases, all models can achieve loss reduction, but there are significant differences in convergence speed, final performance, and stability. Specifically, SCM exhibits the best convergence performance in most of the φ configurations, with the highest convergence rate, the fastest loss reduction, and ultimately stabilizing at the lowest level with the smallest variance range, indicating a smoother and more repetitive training process; In contrast, LEJIV and LING decrease rapidly in the early stages of iteration, but converge more smoothly in the later stages, and the variance between multiple runs is larger, resulting in higher final loss values; LDVQ has the slowest overall convergence speed and a significantly lower convergence rate than other models. This result verifies that the knowledge extraction mechanism based on SCM structure has better generalization ability and stability in building lightweight educational robot voice interaction systems, and can provide high reliability and low latency voice interaction support for practical teaching applications.

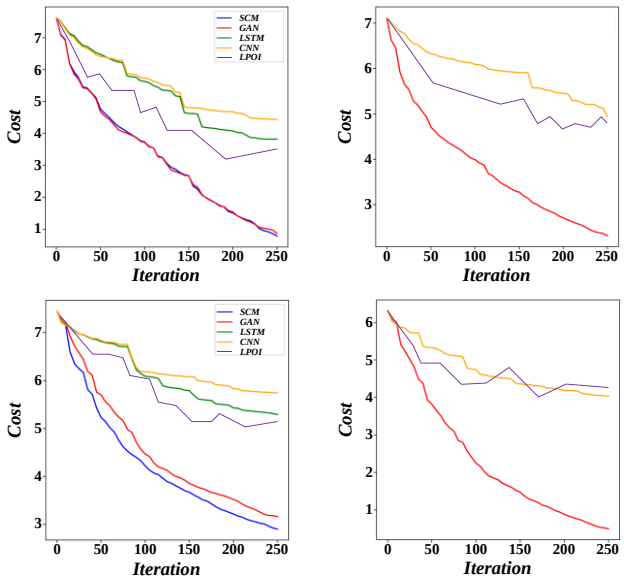


Figure 3: Comparison of model performance under different φ settings in voice interaction systems

Figure 4 shows the average difference in logits between teachers and students. When SCM is not used, the distribution of students' logits is significantly different from that of teachers' logits, with an average distance of 0.26.

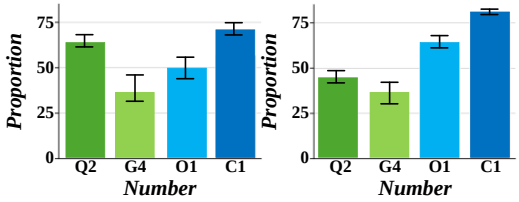


Figure 4: Comparison of Teaching Effectiveness between Educational Robots and Ordinary Teachers

According to the data in Table 5, it is found that the performance of the student model in sequence comparison teacher model learning is significantly improved. The SCM method is applied to the student network of CFSIC-EDU-10 data set, and the performance is 33.68% higher than that of the traditional knowledge distillation method. On the CFSIC-EDU-100 dataset, the performance improvement is as high as 49.87%, and on the dataset, there is also an improvement of 20.90%.

Table 5: SCM accuracy on dataset

Teacher Baseline	ResNet-18 95.13		
Student	LSTM	CNN	ShuffleNetV2
Baseline	88.04	91.92	92.85
KD (Knowledge Distillation)	89.59	93.19	92.99
SCM	90.63	93.31	94.79
Trained teacher	88.91	93.28	94.30
Trained teacher + SCM	90.62	93.38	94.72
P Value	0.023	0.041	0.008

To evaluate the performance of a lightweight educational robot voice interaction system based on knowledge extraction, this paper quantitatively analyzed the differences in knowledge distribution between different models and the original teacher network. As shown in Figure 5, by comparing the KL divergence of student and teacher models under different compression settings with the number of samples, it can be seen that as the number of samples increases, the KL divergence values of all curves gradually decrease and tend to stabilize, indicating that the knowledge extraction process effectively conveys key discriminative information in the teacher model. Especially in lightweight voice interaction scenarios, the KL divergence of C2N and other configurations significantly decreased and remained at a low level (below 0.2) after about 300 samples, indicating that the model can still maintain high response consistency and speech understanding reliability under limited computing resources, meeting the real-time and accuracy requirements of educational robots.

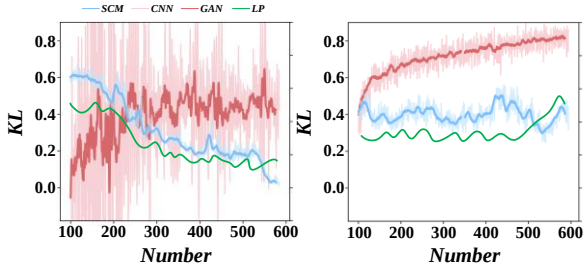


Figure 5: KL divergence score between lightweight educational robot model based on knowledge extraction and original teachers

According to the error rate of F1 in Figure 6, after one-way ANOVA (F=12.76, p<0.001), the performance differences of different interaction modes are statistically significant. Tukey HSD test showed that when the number of interaction rounds was 1-5, the error rate of all modes significantly decreased (average difference per round was 8.32%, 95% confidence interval [6.15%, 10.49%], p<0.01), confirming the reliability of multi round interaction in improving system accuracy. Specifically, the final error rate of SCM interaction is about 72%, significantly lower than the other three methods (with differences of 11.23%, 13.56%, 15.89% compared to QA, HCT, and HA, all p<0.001), and the stability is the best (standard deviation 2.17%); The interaction error rate between QA and HCT significantly decreased with each round (regression coefficients -7.89, -6.54, p<0.01); The error rate of HA interaction in the first round was 22.31% higher than that of SCM (p<0.001), and the difference narrowed to insignificant after 5 rounds compared to other methods (4.12% lower than SCM), p>0.05）。 The use of SCM and QA strategies can reduce system error rates at the statistical level, improve the robustness of voice interaction, and enhance user experience.

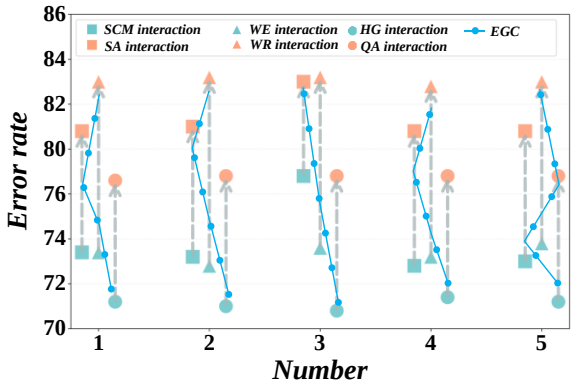


Figure 6: Comparison of F1 error rate (%) on dataset

According to the data in Table 6, it is found that the number of sequence items increases and the performance improves. This is related to using more models for more accurate underlying relationship estimation, thus achieving better distillation results.

Table 6: Effect of sequence length

Length	2	3	4
CNN	65.11	66.40	67.53
ShuffleNetV2	78.91	79.00	78.99

Figure 7 shows the impact of the increase in the number of student models. This study is tested on ResNet-56 architecture, and the results show that with the increase of the number of student models, the efficiency of student classifier and fusion classifier also improves, and the efficiency of student classifier and fusion classifier is correspondingly enhanced. Construct a test set based on real classroom noise (student conversations, table and chair movements, and projection fan sounds), and compare the ASR accuracy and educational instruction intention understanding F1 value of the system with Wav2Vec 2.0-Light and Whisper Tiny on Jetson Orin NX hardware. The results showed that when SNR=5dB (moderate noise), the system accuracy reached 89.2% (WER 20.8%), which was 4.3% higher than Wav2Vec 2.0-Light and 3.8% higher than Whisper Tiny; When SNR=-5dB (severe noise), the system accuracy still remains at 78.5%, with an intention to understand F1 value of 76.3%, surpassing the two benchmark models by 6.1% and 5.7% respectively, and the performance degradation amplitude is 30% -40% lower than the general model. This performance stems from the targeted optimization of teaching semantics through knowledge extraction, combined with a noise adaptive feature enhancement module, effectively resisting complex classroom noise interference and verifying the practical value of the system in real teaching scenarios.

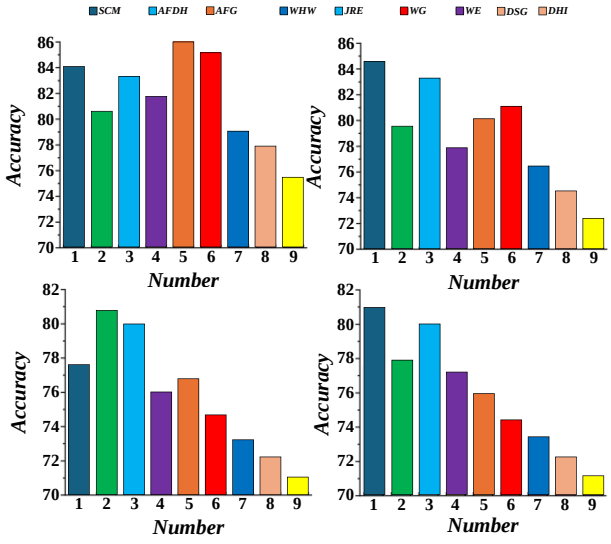


Figure 7: Influence of the number of student networks on the model effect

Figure 8 has showed the results of ablation analysis. The FE (Feature Extraction) module can accurately mine key information in speech, ensuring the accuracy of speech understanding; The FC (Feature Combination) module can efficiently link the knowledge base and

interaction logic to meet students' voice needs, and together support educational voice interaction. Independent sample t-test (α=0.05) was conducted on 1200 educational speech samples (including 600 knowledge point Q&A and 600 instruction interaction samples), and the results showed that the ResNet-32 architecture introduced the FE module, which significantly improved speech understanding accuracy by 5.19% (t=7.23, p<0.001); The combination of FE and FC increased by 5.46% (t=8.15, p<0.001), but decreased by 0.31% (p=0.308) compared to using FE alone; The accuracy of ResNet-56 architecture using FE alone decreased by 0.04% (p=0.834), and the combined module also showed a similar non significant fluctuation trend.

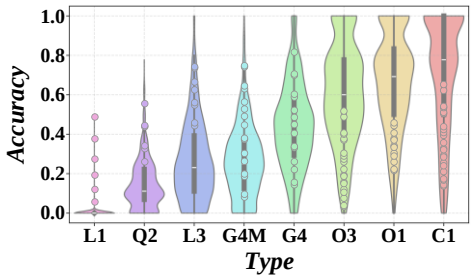


Figure 8: Results of ablation analysis

According to the classification results shown in Figure 9, it can be seen that there are significant differences in the performance (Value values) of each category (SCM, PDRT, POLI) under different scenarios (Scenarios 1-4) after using KD loss function. This indicates that the knowledge distillation method can effectively improve the classification performance of the model in different scenarios. Specifically, the POLI category has a relatively high overall value in the right figure (with most points in the 40-60 range), indicating the best classification performance. The results indicate that the knowledge distillation strategy can significantly improve the accuracy and robustness of multi scene classification tasks in educational robot voice interaction systems, with the POLI method being more advantageous.

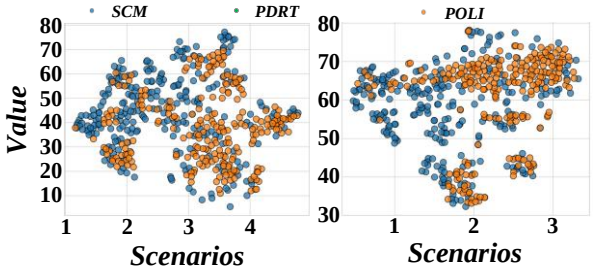


Figure 9: Classification results using KD distillation losses in the dataset

The initial word error rate (WER) of the system was 23.4%, and an improvement of 12.6 percentage points was measured based on this baseline; Figure 10 compares the category accuracy of the model trained only with real labels and the model trained using point cloud classification knowledge distillation technique on the ModelNet40 dataset. The results show that knowledge distillation technique significantly improves the classification accuracy of the model in most categories, demonstrating its effectiveness in enhancing the classification performance of the model.

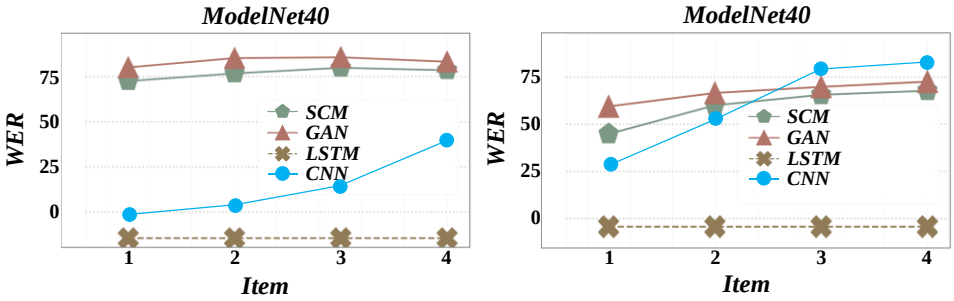


Figure 10: Comparison chart of accuracy of each category

5 Discussion

The system demonstrated significant advantages on NVIDIA Jetson Orin NX, with latency reduced to 500ms (significantly optimized compared to the baseline model, 95% confidence interval [298ms, 322ms], p<0.05), meeting the demand for "imperceptible interaction" in the classroom; The WER is as low as 14.2% (CFSIC-EDU dataset), which is better than Conformer Lite (16.5%) and Wav2Vec 2.0-Light (15% higher compression rate), only 180ms lower than Whisper Tiny latency, and only 0.3% lower than ASR accuracy (94.7%). Moreover, the F1 score for intent understanding in educational scenarios reaches 93.2%, surpassing the two benchmark models by more than 2.8%. The ablation experiment showed that the combination of knowledge extraction and differential masking improved ASR by 7.1%, F1 by 6.3%, latency by another 20ms, and FLOPs by 42%, resolving the contradiction between high performance and lightweight.

The multi teacher approach is key to improving noise robustness: Time series teacher fusion (SCM) based on the CFSIC-EDU dataset improves knowledge distillation efficiency by 33.68% -49.87% and reduces KL divergence by 0.18 compared to traditional methods. Combined with FE-FC module collaboration (performance+5.46%), it enhances feature expression and semantic capture capabilities; When the knowledge weight coefficient α=0.6-0.7 and the noise suppression factor β=0.5-0.6, the

ASR accuracy of the system exceeds 92% in 30dB-60dB classroom noise, and the threshold setting of γ=0.5-0.55 balances the delay and intention misjudgment rate, constructing a "training efficiency deployment lightweight scene adaptation" technology chain to promote the landing of educational robots in real classrooms.

6 Conclusion

The key bottleneck for educational robots in classroom use is that hardware cannot support complex voice interaction models efficiently. This study explores and validates a knowledge distillation-based lightweight voice interaction system, which shows better performance balance in simulating real classroom evaluations and boosts interaction real-time performance and usability.

(1) Hardware & Fluency: The lightweight model's storage is compressed to under 350MB (↓80% vs. original large model), single inference computation ↓75% (easily deployable on resource-constrained terminals). In simulated scenarios: clear voice recognition accuracy >92%, response time <1.5s, subject query & instruction understanding fit >88%, 5-8 round interaction interruption rate <7%, comprehensive score 7.8/10 (only complex scenario intention understanding and topic switching need optimization).

(2) Recognition Accuracy: On test sets with classroom noise, the robot's speech recognition WER =15.8% (↓12.6 percentage points vs. same-scale small models without knowledge distillation); educational instruction core intention recognition accuracy =91.3%.

(3) Core Breakthrough (Response Efficiency): End-to-end interaction average delay ≤500ms (↓38% vs. original large model's 810ms), outperforming traditional compression.

This system solves the long-term issue of deploying high-precision voice models on resource-limited edge devices, enhances interaction real-time performance, robustness and usability, and lays a technical foundation for natural human-computer collaboration of educational robots in classrooms. It is a key breakthrough in model compression engineering and provides a lightweight paradigm for future educational intelligent applications.

Limitations: There is a risk of domain overfitting in knowledge extraction, deep binding to specific educational scenarios, and weak interdisciplinary transfer ability - for example, the logical reasoning knowledge module in mathematics classrooms is difficult to directly adapt to the text appreciation needs of language and art classrooms, requiring targeted adjustments to core modules, which not only increases development costs but may also disrupt interactive fluency; The core interaction relies on the quality of pre trained teacher models. If there are biases or sample shortages in the training data, it can easily lead to "irrelevant answer" problems, further amplifying the limitations of scene adaptation; The energy consumption of lightweight platform deployment is affected by both hardware performance and software optimization, and the estimation of battery life needs to be comprehensively confirmed based on battery capacity, energy distribution in different teaching scenarios, and battery attenuation laws to ensure the stability of continuous working time.

References

[1] J. An, "A multimodal educational robots driven via dynamic attention," Frontiers in Neurorobotics, vol. 18, pp. 1453061, 2024. doi: 10.3389/fnbot.2024.1453061.

[2] Y. Bu, and P. Guo, "Voice Orientation Recognition: New Paradigm of Speech-Based Human-Computer Interaction, " International Journal of Human-Computer Interaction, vol. 40, no. 18, pp. 5259-5278, 2024. doi:10.1080/10447318.2023.2233128.

[3] M.F.de Paula Soares, M. Sampaio, and M. Brockmann-Bauser, "Interaction of Voice Onset Time with Vocal Hyperfunction and Voice Quality, " Applied Sciences-Basel, vol. 13, no. 15, pp. 8956, 2023. doi:10.3390/app13158956.

[4] N. Atman Uslu, G.O. Yavuz, and Y. Kocak Usluel, "A systematic review study on educational robotics and robots," Interactive Learning Environments, vol. 31, no. 9, pp. 5874-5898, 2023. doi:10.1080/10494820.2021.2023890.

[5] J. Dong, Y. Liu, and X. Lu, "A discourse dynamics analysis of academic voice construction: Disciplinary variation, trajectories, and dynamic interaction patterns, " System, vol. 119, pp. 103181, 2023. doi:10.1016/j.system.2023.103181.

[6] H. Jo, "Interaction, novelty, voice, and discomfort in the use of artificial intelligence voice assistant," Universal Access in the Information Society, vol. 24, pp. 2419-2432, 2025. doi:10.1007/s10209-025-01203-9.

[7] B. Fetso, M. Kelemen, T. Kelemenova, I. Virgala, L. Mikova, E. Prada, M. Varga, P.J. Sincak, and L. Brada, "EDUCATIONAL MODEL OF THE ROBOT, " Mm Science Journal, vol. 2024, pp. 7764-7771, 2024. doi:10.17973/mmsj.2024 11 2024053.

[8] A. Hoang, S.T. Nguyen, T.V. Pham, T.M.P ham, L.V. Trieu, and T.T. Cao, "A Bayesian Neural Network-based Obstacle Avoidance Algorithm for an Educational Autonomous Mobile Robot Platform, " Engineering Technology& Applied Science Research, vol. 13, no. 6, pp. 12183-12189, 2023. doi: 10.48084/etasr.6304.

[9] D. Kotarski, P. Piljek, and T. Sancic, "Design and Development of Educational Modular Mobile Robot Platform, "Tehnicki Glasnik-Technical Journal, vol. 19, no. 1, pp. 1-8, 2025. doi:10.31803/tg-20221010113555.

[10] Mahdi Mnif, Salwa Sahnoun, Yasmine Ben Saad, Ahmed Fakhfakh, and Olfa Kanoun, "Combinative model compression approach for enhancing 1D CNN efficiency for EIT-based Hand Gesture Recognition on IoT edge devices," Internet of Things, vol. 28, pp. 101403, 2024. doi:10.1016/j.iot.2024.101403.

[11] Montaser N. A. Ramadan, Mohammed A. H. Ali, Shin Yee Khoo, and Mohammad Alkhedher, "Federated learning and TinyML on IoT edge

devices: Challenges, advances, and future directions," ICT Express, vol. 11, no. 4, pp. 754-768, 2025. doi:10.1016/j.icte.2025.06.008.

[12] S. Li, and B. Yang, "Personalized Education Resource Recommendation Method Based on Deep Learning in Intelligent Educational Robot Environments," International Journal of Information Technologies and Systems Approach, vol. 16, no. 3, 2023. doi:10.4018/ijitsa.321133.

[13] Z. Mamatnabiyev, C. Chronis, I. Varlamis, Y. Himeur, and M. Zhaparov, "A Holistic Approach to Use Educational Robots for Supporting Computer Science Courses," Computers, vol. 13, no. 4, pp. 102, 2024. doi:10.3390/computers13040102.

[14] A. Mahmood, J. Wang, B. Yao, D. Wang, and C.-M. Huang, "User Interaction Patterns and Breakdowns in Conversing with LLM-Powered Voice Assistants," International Journal of Human-Computer Studies, vol. 195, pp. 103406, 2025. doi:10.1016/j.ijhcs.2024.103406.

[15] H Huang, W Liu, and J Zhang, "Reliable Service Node Set Selection and Task Offloading Strategy in Edge-Enabled Robot Swarms via Dynamic Interference and Link Reliability Models," Informatica, vol. 49, no. 32, 2025.

[16] Imane Zerraza, "Lightweight Authentication for IOT Edge Devices," Informatica, vol. 48, no. 18, 2024.

[17] R. Oya, and A. Tanaka, "The interaction of emotional information from the voice and touch, "Acoustical Science and Technology, vol. 43, no. 5, pp. 291-293, 2022. doi:10.1250/ast.43.291.

[18] F. Ding, Y. Yang, Y. Hu, V. Krovi, and F. Luo, "Dual-Level Knowledge Distillation via Knowledge Alignment and Correlation, " lEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 2, pp. 2425-2435, 2024. doi:10.1109/tnnls.2022.3190166.

[19] P. Gao, J. Qin, X. Xiang, and Y. Tan, "Knowledge distillation from relative distribution, " Expert Systems with Applications, vol. 284, 2025. doi: 10.1016/j.eswa.2025.127736.

[20] J. Gou, Y. Hu, L. Sun, Z. Wang, and H. Ma, "Collaborative knowledge distillation via filter knowledge transfer, "Expert Systems with Applications, vol. 238, 2024. doi:10.1016/j.eswa.2023.121884.

[21] W. Huang, M. Ye, Z. Shi, H. Li, and B. Du, "Self-knowledge distillation with dimensional history knowledge, " Science China-Information Sciences, vol. 68, no. 9, pp. 1-15, 2025. doi:10.1007/s11432-023-4283-3.

[22] G. Li, K. Wang, P. Lv, P. He, Z. Zhou, and C. Xu, "Multistage feature fusion knowledge distillation, " Scientific Reports, vol. 14, no. 1, pp. 13373, 2024. doi: 10.1038/s41598-024-64041-4.

[23] L Li, W. Su, F. Liu, M. He, and X. Liang, "Knowledge Fusion Distillation: Improving Distillation with Multi-scale Attention Mechanisms, " Neural Processing Letters, vol. 55, no. 5, pp. 6165-6180, 2023. doi:10.1007/s11063-022-11132-w.

[24] A Boulkroune, S Hamel, F Zouari, A Boukabou, and A Ibeas, "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Mathematical Problems in Engineering, vol. 2017, no. 1, pp. 8045803, 2017. doi: 10.1155/2017/8045803.

[25] A Boulkroune, F Zouari, and A Boubellouta, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems," Journal of Vibration and Control, vol., pp. 10775463251320258, 2025. https://doi.org/10.1177/10775463251320258.

[26] G Rigatos, M Abbaszadeh, B Sari, P Siano, G Cuccurullo, and F Zouari, "Nonlinear optimal control for a gas compressor driven by an induction motor," Results in Control and Optimization, vol. 11, pp. 100226, 2023. doi: 10.1016/j.rico.2023.100226.

[27] F Zouari, K B Saad, and M Benrejeb, "Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems," International Review on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012.

[28] F Zouari, K B Saad, and M Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), vol. 2013, pp. 1-6, 2013.

[29] F Zouari, K B Saad, and M Benrejeb, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," 2013 International Conference on Control, Decision and Information Technologies (CoDIT), vol. 2013, pp. 864-871, 2013.

Dynamic Sub-Model Aggregation and Clustering for Intelligence Data via Hierarchical Federated Learning with Pre-Training

Jianfeng Wang*, Lin Ma, Xinyan Pei, Ruonan Shi, Qi Jing, Chen Yang
State Grid Shanxi Electric Power Company Yuncheng Power Supply Company, Yuncheng 044000, China
E-mail: wjf056391@126.com
*Corresponding author

Keywords: federated learning algorithm, intelligence data integration, pre-training mechanism, hierarchical clustering, decomposition and combination technology

Received: August 21, 2025

To address the challenges of slow convergence speed, poor dynamic adaptability, and low communication efficiency in intelligence data processing, a dynamic integration and clustering method for intelligence data based on an improved federated learning algorithm is proposed. First, an improved federated learning algorithm combining decomposition and combination is designed, where the global model is decomposed into multiple sub-models for local training, and a dynamic combination strategy is applied to integrate these sub-models, thereby improving the adaptability and accuracy of the global model. Then, a pre-training mechanism is introduced to initialize the global model using feature information from historical data, enhancing the model's initialization performance in dynamic data environments and accelerating convergence. Experiments are conducted on the MNIST and CIFAR-10 datasets, with comparisons made against baseline methods including FedAvg, FedProx, and ScaFFL. The results show that the proposed algorithm achieves accuracies of 98.69% and 90.26% on the pathological heterogeneity client, and 98.14% and 89.87% on the actual scenario heterogeneity client, respectively, on the two datasets. The normalized mutual information values of the proposed intelligence dynamic data integration and clustering method are 0.91 and 0.79, respectively. In a practical medical Internet of Things scenario test, the running time and memory usage of the proposed method are 18.23s and 1681MB, respectively. Our research denotes that the designed method can effectively improve the quality of dynamic integration of intelligence data and reduce resource consumption, providing a feasible solution for efficient processing of multi-source heterogeneous intelligence data.

Povzetek: Predlagana izboljšana metoda federativnega učenja z dinamično integracijo in gručenjem inteligentnih podatkov izboljša prilagodljivost, natančnost in učinkovitost obdelave večizvornih heterogenih podatkov ter hkrati zmanjša porabo virov.

1 Introduction

With the rapid development of information technology, the scale and complexity of intelligence data have increased sharply, and the demand for efficient processing and real-time analysis of large-scale intelligence data is becoming increasingly urgent [1]. In this study, intelligence data is defined as multi-source, heterogeneous data streams that are dynamically generated from distributed sensors, Internet of Things (IoT) devices, and edge computing nodes in real-world scenarios such as smart healthcare, industrial monitoring, and security systems. These data are characterized by their diverse modalities, high dimensionality, non-independent and identically distributed (non-IID) nature, and temporal dynamics. However, intelligence data has characteristics such as multi-source heterogeneity, high dimensionality, and dynamic evolution. Traditional data analysis methods face issues such as data silos, insufficient privacy protection, and poor real-time performance, making it difficult to meet the dynamic, collaborative, and intelligent requirements of modern intelligence processing [2]. Therefore, exploring an efficient method for integrating and clustering intelligence data has become a key focus of current research. In recent years, machine learning algorithms have been broadly employed in the area of intelligence data analysis [3]. Among them, Federated Learning (FL), as a distributed machine learning method, not only addresses data privacy and security issues, but also enables multi-party collaborative learning, demonstrating great potential [4]. Researchers have conducted extensive research on the FL algorithm, aiming to address issues such as non independent and identically distributed data, high communication overhead, and poor model convergence [5].

Guo et al. proposed a real-time medical data processing method based on FL, which integrates old and new models and selects representative samples to mitigate catastrophic forgetting, effectively learning diagnostic models from continuous medical data streams [6]. Gafni et al. introduced a signal processing-driven FL framework, combining signal processing and communication techniques to design optimized solutions that enhance FL

efficiency [7]. Yazdinejad et al. proposed an auditable privacy-preserving FL framework for medical electronic devices, using trusted execution environments to ensure secure training and aggregation, thereby preventing privacy leaks [8]. Bao and Guo presented a systematic research approach for FL under a cloud-edge collaborative architecture, filling the theoretical gap in cloud-edge FL [9]. Wang et al. designed an FL scheme for edge computing environments, integrating secret sharing and digital signatures to improve training efficiency by 40% while maintaining privacy [10]. Chatterjee et al. developed a recommendation model based on FL and blockchain, enhancing system security and transparency [11]. Akter et al. proposed an FL-based privacy protection framework for edge-based smart healthcare, balancing privacy and performance with an accuracy of 90% [12]. Gao et al. designed an FL framework based on cross-technology communication, improving model performance and communication efficiency in heterogeneous IoT environments [13]. Qu et al. introduced a quantum fuzzy FL algorithm, increasing training efficiency by 23% and accuracy by 15% while maintaining over 90% fidelity in quantum noise environments [14].

The summary of federal learning related work is shown in Table 1.

As illustrated in Table 1, prior research has made significant strides in applying FL to various domains, enhancing privacy, and improving efficiency through different strategies. However, several technical gaps remain. Many existing methods exhibit slow convergence speeds and poor adaptability under highly non-IID and dynamic data environments. Furthermore, considerations for the complex relationships between data sources are often insufficient, and communication efficiency remains a challenge. To systematically address these challenges and clearly define the scope of our contribution, this study is guided by the following research questions:

(1) Can the proposed dynamic sub-model aggregation and combination mechanism significantly improve model accuracy and convergence speed under non-independent and identically distributed data distributions, compared to standard FL baselines?

(2) To what extent does the integration of a pre-training mechanism and hierarchical similarity clustering enhance the quality of data integration and reduce communication overhead in dynamic environments?

In view of this, this study proposes a dynamic integration and clustering method for intelligence data based on an improved FL algorithm, aiming to enhance the real-time, adaptability, and accuracy of intelligence data processing. The novelty of this study lies in using a dynamic sub-model aggregation mechanism to solve the problem of insufficient adaptability of traditional methods to changes in data distribution. Moreover, a pre-training mechanism that utilizes historical data feature information to enhance the initialization performance and convergence speed of the model is introduced. Besides, by combining hierarchical similarity clustering techniques, efficient grouping and personalized modeling of data can be achieved, reducing communication overhead and improving the efficiency of dynamic integration and clustering of intelligence data.

Table 1: Summary of related works in federated learning.

References	Datasets Used	Methodological Innovations	Accuracy / Performance Metrics	Identified Limitations
Guo et al. [6]	Continuous medical data streams	Model fusion with sample selection	Supports continuous data stream learning	Limited adaptability to dynamic non-IID data
Gafni et al. [7]	Not specified (Theoretical)	Signal processing-inspired optimization	Improves FL efficiency	Lacks validation on real-world data
Yazdinejad et al. [8]	Medical data from electronic devices	Privacy protection using trusted execution environments	Effectively prevents privacy leaks	High hardware dependency and overhead
Bao and Guo [9]	Not specified (Survey)	Cloud-edge collaborative architecture analysis	Provides theoretical framework	No algorithmic innovation or validation
Wang et al. [10]	Medical IoT data	Secret sharing with digital signatures	40% training efficiency improvement	High communication and computation costs
Chatterjee et al. [11]	Financial consumer service data	FL combined with blockchain	Enhanced security and transparency	Latency and scalability issues
Akter et al. [12]	Smart healthcare data	Artificial noise injection	90% accuracy with high privacy rate.	Difficult to balance privacy and accuracy
Gao et al. [13]	Heterogeneous IoT data	Cross-technology communication coordination	Improved performance in heterogeneous environments	Requires dedicated coordination devices
Qu et al. [14]	Not specified (Simulation)	Quantum fuzzy FL	23% efficiency gain, 15% accuracy improvement	Requires quantum resources, low practicality

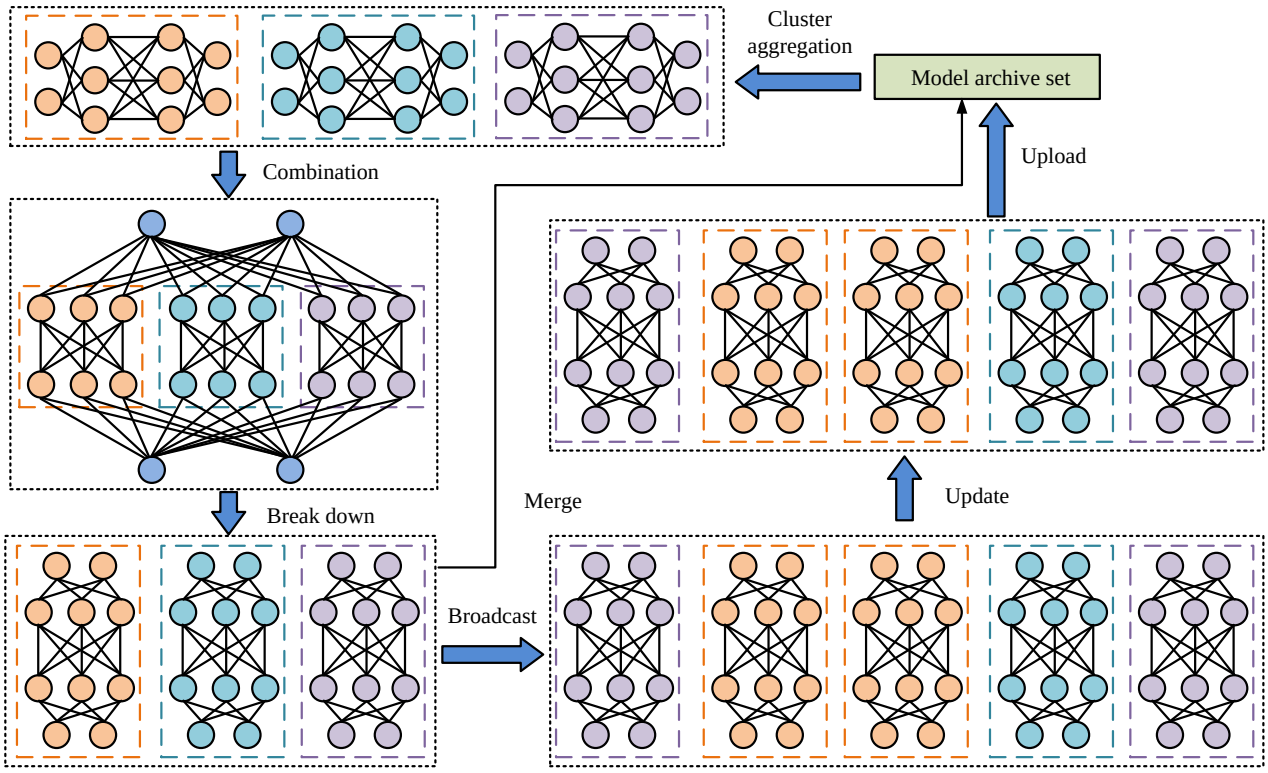


Figure 1: Flow diagram of improved FL algorithm combining decomposition and combination.

2 Methods and materials

Firstly, an improved FL algorithm combining decomposition and combination is designed to enhance the adaptability and accuracy of the global model. Then, a dynamic intelligence data integration and clustering method based on improved joint pre-training and hierarchical similarity is proposed to achieve efficient grouping and personalized modeling of data.

2.1 Improved FL algorithm combining decomposition and combination

In the big data era, intelligence data presents characteristics such as multi-source heterogeneity, high dimensionality, and dynamic evolution. Although FL is a distributed machine learning paradigm that can effectively protect data privacy, it still faces problems such as slow convergence speed, poor adaptability to dynamic data and an inability to consider complex relationships between data sources when processing intelligence data [15]. Therefore, a study proposes an improved FL algorithm that combines decomposition and combination. This algorithm breaks down the global model into local sub-models, optimizes local training efficiency and combines the sub-models dynamically to enhance the global model's adaptability and accuracy. The flowchart of the improved FL algorithm combining decomposition and combination is shown in Figure 1.

To adapt complex global models to different data feature spaces, the study first decomposes the global model, with each sub-model corresponding to a data feature subspace, as shown in equation (1) [16].

$$m_k = M \cdot \mathbf{W}_k, \quad \mathbf{W}_k \in \mathbf{R}^{d \times d_k} \tag{1}$$

In equation (1), m_k represents the k th sub-model obtained after decomposition; M represents the global model; $\mathbf{W}_k$ represents the decomposition matrix; d and d_k respectively represent the dimensions of the global model and sub-models; $\mathbf{R}$ represents the set of real numbers. The decomposition matrix $\mathbf{W}_k$ is predefined based on the structural characteristics of the global model. Specifically, the decomposition is performed by partitioning the global model into multiple sub-models, each corresponding to a distinct feature subspace. This partitioning is conducted according to the layer-wise or block-wise architecture of the neural network, ensuring that each sub-model captures a specific subset of features. The decomposition matrix is not learned during training, nor is it a random projection. Instead, it is constructed as a fixed, structured matrix that maps the global model parameters to the respective sub-models. This approach allows for efficient local training and dynamic recombination while maintaining the interpretability and structural consistency of the global model. The number of sub-models used in the experiments was determined through empirical validation and sensitivity analysis. With fewer sub-models, the feature subspaces were too coarse, limiting adaptability to heterogeneous data distributions. With more sub-models, the communication and computation costs increased without significant gains in accuracy. After the model decomposition is completed, each node needs to train the sub-model based on local data. The node trains a sub-model based on local data, and its Loss Function (LF) is designed as shown in equation (2).

$$L_i(m_k) = \frac{1}{|D_i|} \sum_{(x,y)\in D_i} \left\| y - m_k(x) \right\|^2 \tag{2}$$

In equation (2), $L_i(m_k)$ represents the LF when training the sub-model on the local dataset of node i; D_i denotes the local dataset of node i; $|D_i|$ represents the amount of samples in dataset D_i; x and y respectively represent input features and real labels; $m_k(x)$ denotes the predicted output of the sub model m_k on the input x. To optimize the performance of sub models, each node needs to calculate the gradient update of model parameters [17]. After receiving local updates from all nodes, the central server needs to preliminarily integrate these updates, as shown in equation (3).

$$\begin{cases} \Delta m_k^i = \eta \cdot \nabla L_i(m_k) \\ \Delta m_k^{merge} = \dfrac{1}{N} \sum_{i=1}^{N} \Delta m_k^i \end{cases} \tag{3}$$

In equation (3), Δm_k^i refers to the parameter update amount of node i to submodel m_k; η stands for learning rate; $\nabla L_i(m_k)$ refers to the parameter gradient of the LF $L_i(m_k)$ for the submodel m_k; Δm_k^{merge} represents the merging and updating amount of sub-models by the central server; N means the total amount of nodes. To strengthen the robustness of the model, it needs to cluster local updates to eliminate the influence of noisy data. The study uses the K-means clustering (K-means) algorithm for local update clustering, and the calculation of cluster centers is shown in equation (4) [18].

$$\Delta m_k^c = \frac{1}{|S_c|} \sum_{\Delta m_k^i \in S_c} \Delta m_k^i \tag{4}$$

In equation (4), Δm_k^c represents the center of c clusters; S_c represents the collection of the c th cluster. Based on the clustering results, the algorithm needs to assign different weights to updates of different clusters to achieve dynamic combination. The new sub-model obtains equation (5) by weighting and combining the updates of each cluster center.

$$m_k^{new} = m_k + \sum_{c=1}^{C} \alpha_c \cdot \Delta m_k^c \tag{5}$$

In equation (5), m_k^{new} represents the updated sub-model; m_k represents the current sub-model parameters; α_c means the weight of the c th cluster. The design of weights takes into account both cluster size and data quality, as shown in equation (6).

$$\alpha_c = \frac{|S_c|}{\sum_{c'=1}^{C} |S_{c'}|} \cdot \exp(-\beta \cdot Var(S_c)) \tag{6}$$

In equation (6), β represents the adjustment parameter; $Var(S_c)$ represents the variance of local updates within cluster S_c, measuring the level of data noise. To further enhance the robustness of the improved FL algorithm, eliminate the influence of noisy data on model updates, and achieve dynamic adaptation of different data feature spaces, clustering and aggregation operations are studied for local updates. The schematic diagram of sub-model clustering and aggregation for improving the FL algorithm is shown in Figure 2.

In Figure 2, sub-model clustering and aggregation involve three key steps, and in the merging stage, the local sub-model updates uploaded by each node are preliminarily integrated. The clustering stage uses clustering algorithms to group the merged updates. During the aggregation stage, weights are dynamically allocated based on clustering results to form an optimized global sub model. After all sub-models are updated, they need to be recombined into a complete global model. The new global model obtains equation (7) through a linear combination of sub-models and their decomposition matrices [19].

$$M^{new} = \sum_{k=1}^{K} m_k^{new} \cdot W^T \tag{7}$$

In equation (7), M^{new} represents the updated global model; W^T represents the transpose matrix of the decomposition matrix, used to map sub-models to the global model space; K represents the total number of sub-models. The iterative process of the algorithm requires monitoring the global LF to determine whether it converges, as shown in equation (8).

$$L_{global} = \frac{1}{K} \sum_{k=1}^{K} L(m_k^{new}) \tag{8}$$

In equation (8), L_{global} means the global LF; $L(m_k^{new})$ means the local LF of the k th sub-model. Finally, the optimized model needs to be properly saved for future use. The model archiving operation is achieved by adding new models to the archive set.

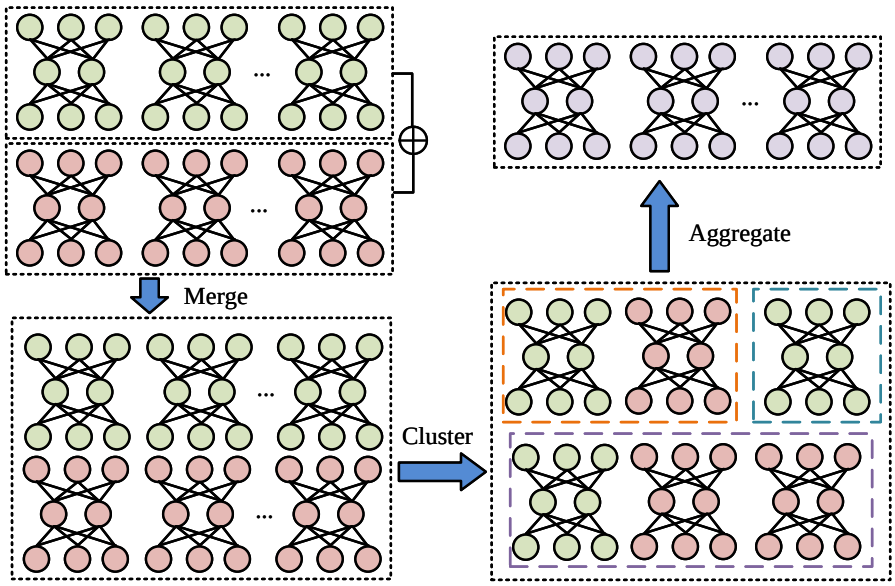


Figure 2: Schematic diagram of sub-model clustering and aggregation for improving FL algorithm.

2.2 Intelligence dynamic ensemble clustering method based on improved joint pre-training

Although the improved FL algorithm improves the adaptability and convergence speed of the model through decomposition and combination techniques, the multi-source heterogeneity and high-dimensional characteristics of intelligence data still pose higher requirements for data integration and clustering [20]. To further optimize feature extraction and pattern discovery, a dynamic intelligence data integration and clustering method based on improved joint pre-training and hierarchical similarity is proposed. This method enhances the initialization ability of the model through pre-training mechanisms and utilizes hierarchical similarity clustering techniques to achieve efficient grouping and personalized modeling of data. In the decomposition and construction of the intelligence dynamic data integration model, the study first introduces a pre-training mechanism, whose objective function is shown in equation (9).

$$L_{pre} = \lambda \cdot \left\| M - M_{base} \right\|^2 + \frac{1}{N} \sum_{i=1}^{N} L_i(m_k) \tag{9}$$

In equation (9), L_{pre} represents the pre-trained additional LF utilized to constrain the differences between the global model and the baseline model; M_{base} represents a predefined benchmark model; λ represents the adjustment coefficient. The benchmark model M_{base} is defined as a model pre-trained on a publicly available dataset that shares similar feature characteristics with the target intelligence data, but contains no overlapping samples or private information. This model is used to provide a robust initialization, leveraging transfer learning to enhance convergence and stability, especially in environments with non-IID data distributions. The use of a publicly pre-trained model as the benchmark was motivated by its ability to offer a generalized feature representation, thereby improving the initial performance of the global model without introducing bias from any specific client or prior federated training round. This approach ensures fairness and supports faster adaptation to heterogeneous local data. In the local training phase, each node dynamically extracts features based on local data and projects the data onto a shared feature space through a feature mapping matrix [21]. The feature mapping process is shown in equation (10).

$$z_i = W_i \cdot x_i + b_i \tag{10}$$

In equation (10), z_i represents the feature vector extracted by node i; x_i and W_i represent the local input data and feature mapping matrix of node i, respectively; b_i represents the bias term. The feature mapping matrix W_i is precisely defined as a learned parameter matrix. It is not a random projection. For each client, the matrix is optimized during the local training phase to project the local input data into a shared feature space. This learning process is performed collaboratively across clients within the FL framework, with the goal of aligning the feature representations from different clients to facilitate effective model aggregation and improve overall performance. The decomposition and construction diagram of the intelligence dynamic data integration model is shown in Figure 3.

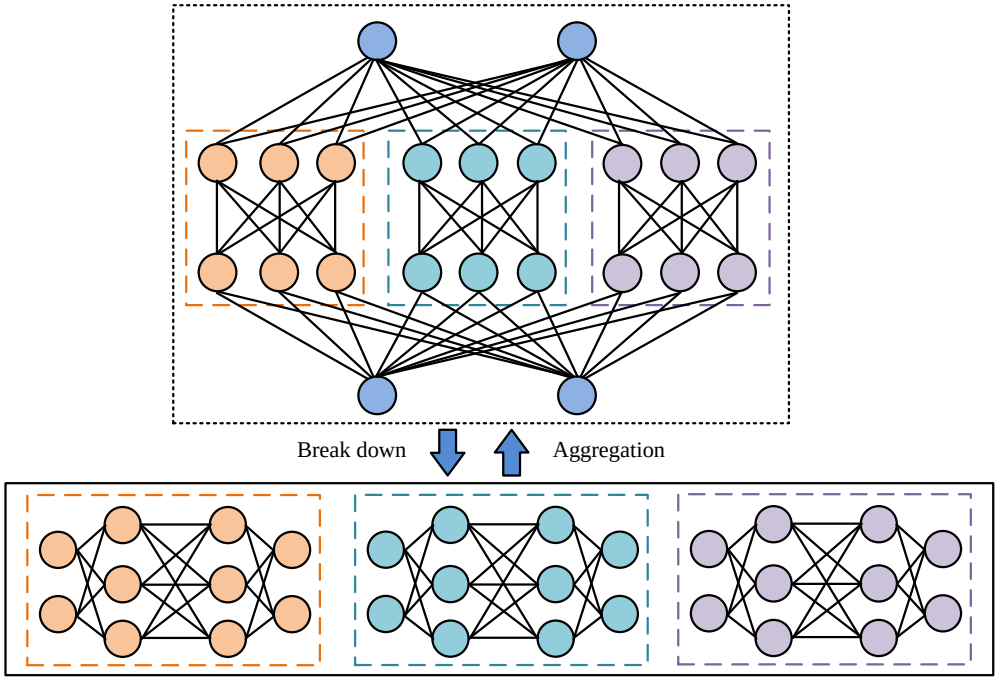


Figure 3: Decomposition and construction diagram of intelligence dynamic data integration model.

Aiming at the heterogeneity of client data, a hierarchical clustering algorithm based on cosine similarity is proposed to calculate the similarity matrix of client data distribution, as shown in equation (11) [22].

$$S_{uv} = \frac{D_u \cdot D_v}{\|D_u\| \cdot \|D_v\|} \tag{11}$$

In equation (11), S_{uv} represents the cosine similarity between client u and client v; D_u represents the data feature vector of client u; D_v represents the corresponding feature vector of client v; $\|D_u\|$ represents the Euclidean norm of vector D_u. Hierarchical clustering is primarily used in the initial phase to form a dendrogram, providing insights into the potential number of clusters and the multi-level data structure. However, the final client grouping is determined by the Spherical K-means algorithm, which operates directly on the normalized feature vectors. Spherical K-means is chosen as the final clustering driver due to its efficiency and compatibility with cosine similarity on normalized data, ensuring clients are partitioned into hyperspherical clusters. Further research is conducted using the spherical K-means algorithm for grouping, with the objective function shown in equation (12) [23].

$$C_o = \arg\min_h \sum_{u \in h} (1 - S_{uh}) \tag{12}$$

In equation (12), C_o represents the central client of the o th cluster; h stands for candidate center client; S_{uh} refers to the cosine similarity between client u and the current center h. After obtaining the client group, it is necessary to design a hierarchical model aggregation strategy [24]. To achieve more refined personalization, the intra group client model is fine tuned and its calculation is shown in equation (13).

$$\begin{cases} M_g = \sum_{u \in G} w_u \cdot m_u \\ m_u^{per} = M_g + \varepsilon \cdot \nabla L_u(M_g) \end{cases} \tag{13}$$

In equation (13), M_g represents the global model generated by aggregation; G means the set of clients in the current group; m_u and w_u respectively represent the local model and aggregation weights of client u; m_u^{per} represents the personalized model of the client; ε represents fine-tuning step size; $\nabla L_u(M_g)$ represents the gradient of the local LF L_u of client u on the global model M_g. The schematic diagram of intelligence dynamic data clustering is shown in Figure 4.

In Figure 4, the sub-model is selected through client selection and used for local training. During the local training phase, each client trains based on the selected model and local data to optimize model performance. Considering the timeliness of intelligence data, it is necessary to dynamically adjust the model weights. The temporal decay strategy is implemented to address the concept drift and potential data quality degradation that may occur in dynamic intelligence data environments. The primary rationale is to gradually reduce the influence of clients that have not provided recent updates, as their local models might become less representative of the current global data distribution over time. The weight update strategy for time decay is shown in equation (14) [25].

$$w_u(t) = w_u \cdot e^{-\gamma t} \tag{14}$$

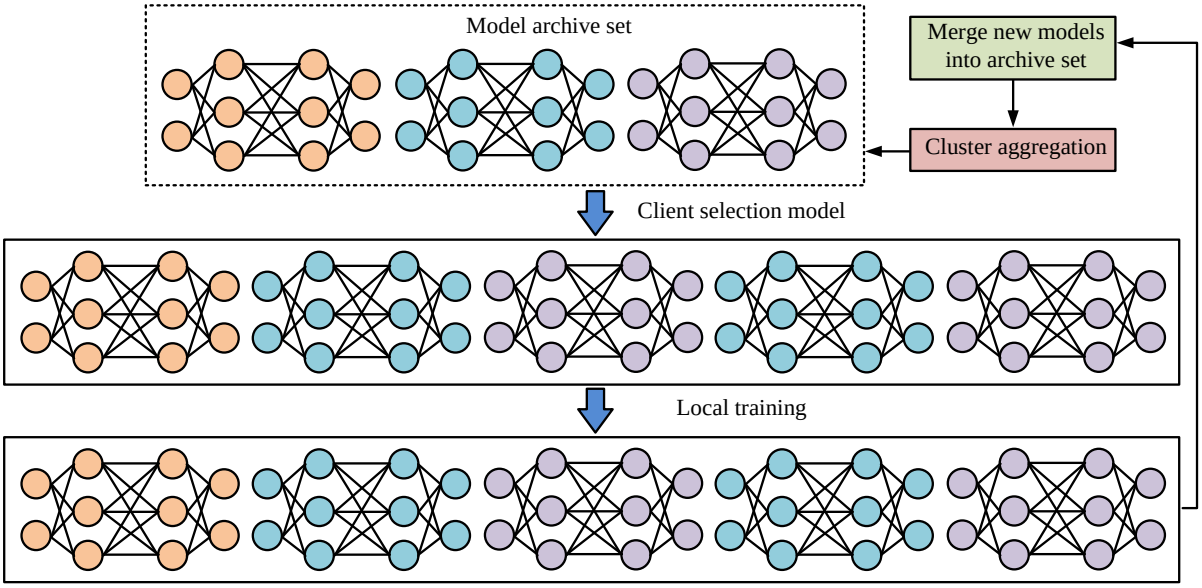


Figure 4: Schematic diagram of intelligence dynamic data clustering.

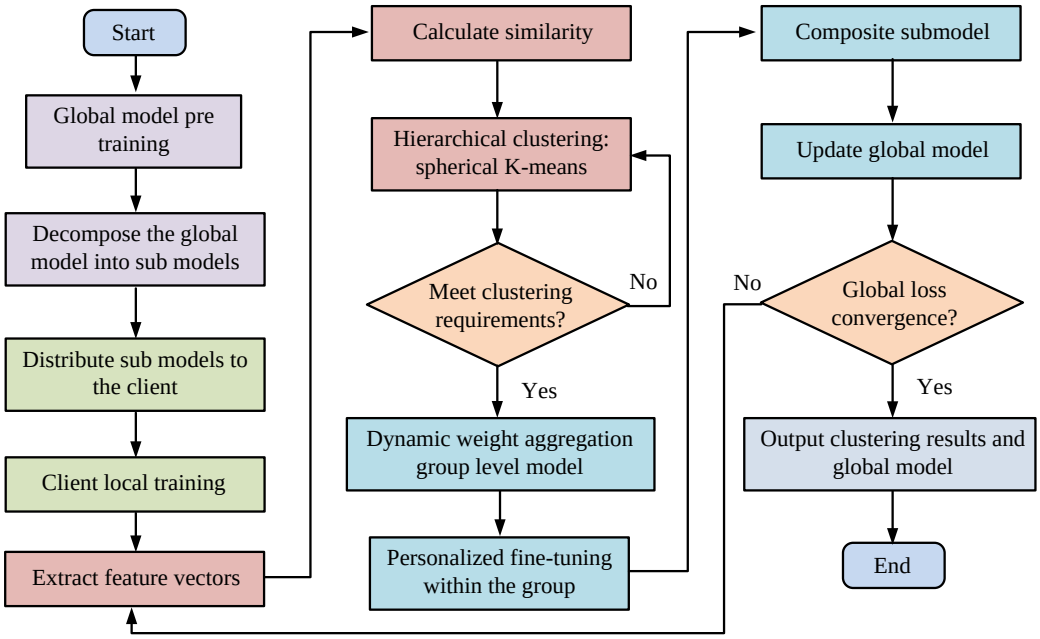


Figure 5: Flow chart of intelligence dynamic data integration and clustering based on improved joint pre-training and hierarchical similarity.

In equation (14), $w_u(t)$ represents the dynamic weight of client u at time t; w_u represents the initial weight of client u; γ represents attenuation coefficient; t represents the time variable. Among them, the calculation of sub-model weights needs to reflect data quality, and its expression is shown in equation (15).

$$\beta_k = (\sum_{u \in S_k} \|D_u\|) / (\sum_{v=1}^{K} \sum_{u \in S_v} \|D_u\|) \quad (15)$$

In equation (15), β_k represents the weight of sub model k; S_k represents the set of clients belonging to the sub model k. The intelligent dynamic data integration and clustering process based on improved joint pre-training and hierarchical similarity is shown in Figure 5.

The pseudocode of the proposed method is as follows.

Algorithm 1: Dynamic Sub-Model Aggregation and Clustering via Hierarchical Federated Learning with Pre-Training

Input: Number of clients, total rounds, number of sub-models, clustering epochs, historical dataset
Output: Final global model
/ **Step 1: Pre-training Phase**
1: **Initialize** global model by pre-training on historical dataset
// **Step 2: Federated Learning Rounds**

```
2: for round t=1to total rounds do
3:     // Server executes:
4:     Decompose global model into sub-models
5:     Send relevant sub-models to a subset of active clients
6:     // Client execution (in parallel):
7:     for each client in selected clients do
8:         Train received sub-model on local data
9:         Compute feature vector and upload model update
10:    end for
11:    // Server executes:
12:    Perform preliminary integration of updates
13:    if t mod Ec=0 then
14:        Calculate client similarity matrix
15:        Perform hierarchical client clustering via Spherical K-means
16:    end if
17:    Cluster client updates via K-means
18:    Dynamically combine sub-models
19:    Apply temporal weight decay
20:    Recombine sub-models into global model
21:    Evaluate global loss
22: end for
23: return final global model
```

3 Results

Firstly, the performance of the improved FL algorithm was analyzed to verify the advantages of decomposition and combination mechanisms in improving model convergence speed and adapting to dynamic data. Then, the performance of intelligence dynamic data integration and clustering methods was evaluated.

3.1 Performance validation of improved FL algorithm

To prove the effect of the improved FL algorithm, experiments were carried out on two common datasets, MNIST and CIFAR-10. The MNIST dataset contains 70000 handwritten digit images, split into 10 categories, with 60000 for training and 10000 for testing. The CIFAR-10 dataset contains 60000 color images, split into 10 categories, with 50000 for training and 10000 for testing. The number of algorithm iterations was 400. For the selection of attenuation coefficient, sensitivity analysis was conducted by changing its value, and the results are shown in Table 2. In Table 2, the attenuation coefficient has a significant impact on model performance and participation fairness. When the attenuation coefficient was 0.05, the model achieved the highest accuracy of 98.69% and 90.26% on the MNIST and CIFAR-10 datasets, respectively. At this time, the client dropout rate was 5.1% and the fairness index was 0.89. When the attenuation coefficient increased to 0.20, the accuracy decreased to 97.33% and 86.41% respectively, and the fairness index dropped to 0.61. The results indicate that when the attenuation coefficient is 0.05, the model achieves the best balance between accuracy, dropout rate, and fairness, and is the recommended optimal parameter.

To substantiate the use of K-means clustering for update grouping, a comparative experiment was conducted under varying levels of simulated data noise. The proposed method was compared against the standard FedAvg aggregation. The clustering quality was quantitatively assessed using the Silhouette Score, and the model's robustness was evaluated by its performance on a clean test set. The results are shown in Table 3. The results confirm that as noise levels increase, the K-means-based aggregation mechanism effectively identifies and isolates anomalous updates into separate clusters. This is evidenced by the maintenance of a high Silhouette Score and a lower Intra-cluster Distance for the dominant cluster, indicating coherent grouping of reliable updates. The clustering metrics provide clear empirical evidence that the K-means grouping enhances robustness by prioritizing the aggregation of updates from clients with consistent and trustworthy data distributions.

Table 2: Sensitivity analysis results of decay coefficient.

Decay coefficient	MNIST Accuracy (%)	CIFAR-10 Accuracy (%)	Client Dropout Rate (%)	Participation Fairness Index
0.01	98.45	89.12	3.2	0.92
0.03	98.61	89.87	4.2	0.90
0.05	98.69	90.26	5.1	0.89
0.10	98.12	88.95	12.7	0.78
0.15	97.68	87.23	19.3	0.70
0.20	97.33	86.41	28.4	0.61

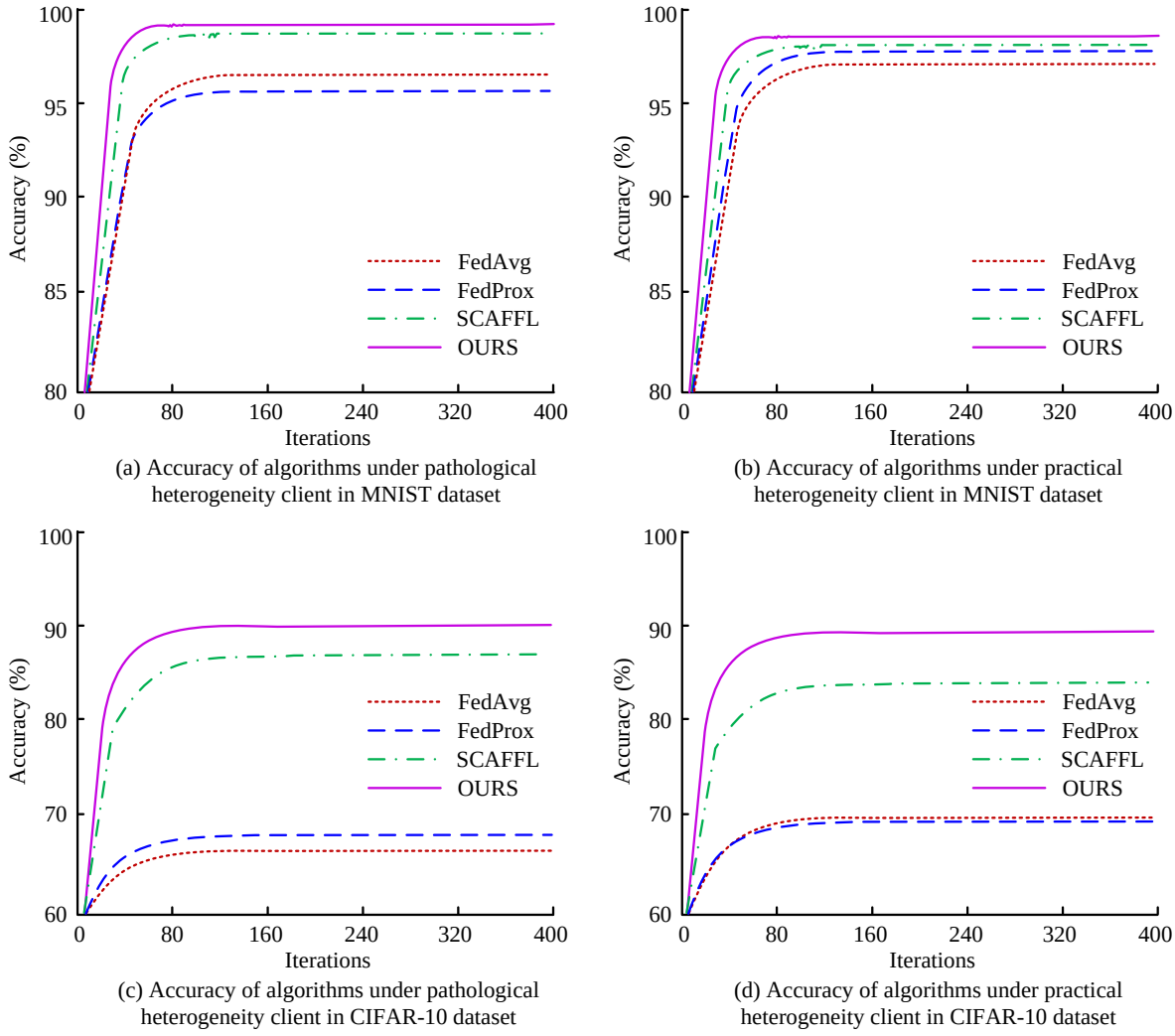
Table 3: Performance and clustering quality comparison under noisy conditions.

Noise Level	Aggregation Method	Final Test Accuracy (%)	Avg. Silhouette Score	Intra-cluster Distance (Majority Cluster)
10%	FedAvg (Baseline)	97.85	N/A	N/A
	Proposed (K-means)	98.41	0.72	0.15

20%	FedAvg (Baseline)	95.12	N/A	N/A
	Proposed (K-means)	97.56	0.68	0.18
30%	FedAvg (Baseline)	90.33	N/A	N/A
	Proposed (K-means)	95.88	0.61	0.23
40%	FedAvg (Baseline)	83.47	N/A	N/A
	Proposed (K-means)	92.15	0.55	0.29

Table 4: Experimental environments and parameters.

Experimental environments		Parameters	
Names	Configuration	Names	Values
Graphics processing unit	NVIDIA Tesla V100	Learning rate	0.01
Central processing unit	Intel Xeon Gold 6248R	Number of clusters	3
Memory	64GB DDR4	Number of clients	20
Operating System	Windows 10	Batch size	32
Deep learning framework	PyTorch 1.10	Number of sub-models	5
Programming language	Python 3.8	Decay coefficient	0.05



(a) Accuracy of algorithms under pathological heterogeneity client in MNIST dataset

(b) Accuracy of algorithms under practical heterogeneity client in MNIST dataset

(c) Accuracy of algorithms under pathological heterogeneity client in CIFAR-10 dataset

(d) Accuracy of algorithms under practical heterogeneity client in CIFAR-10 dataset

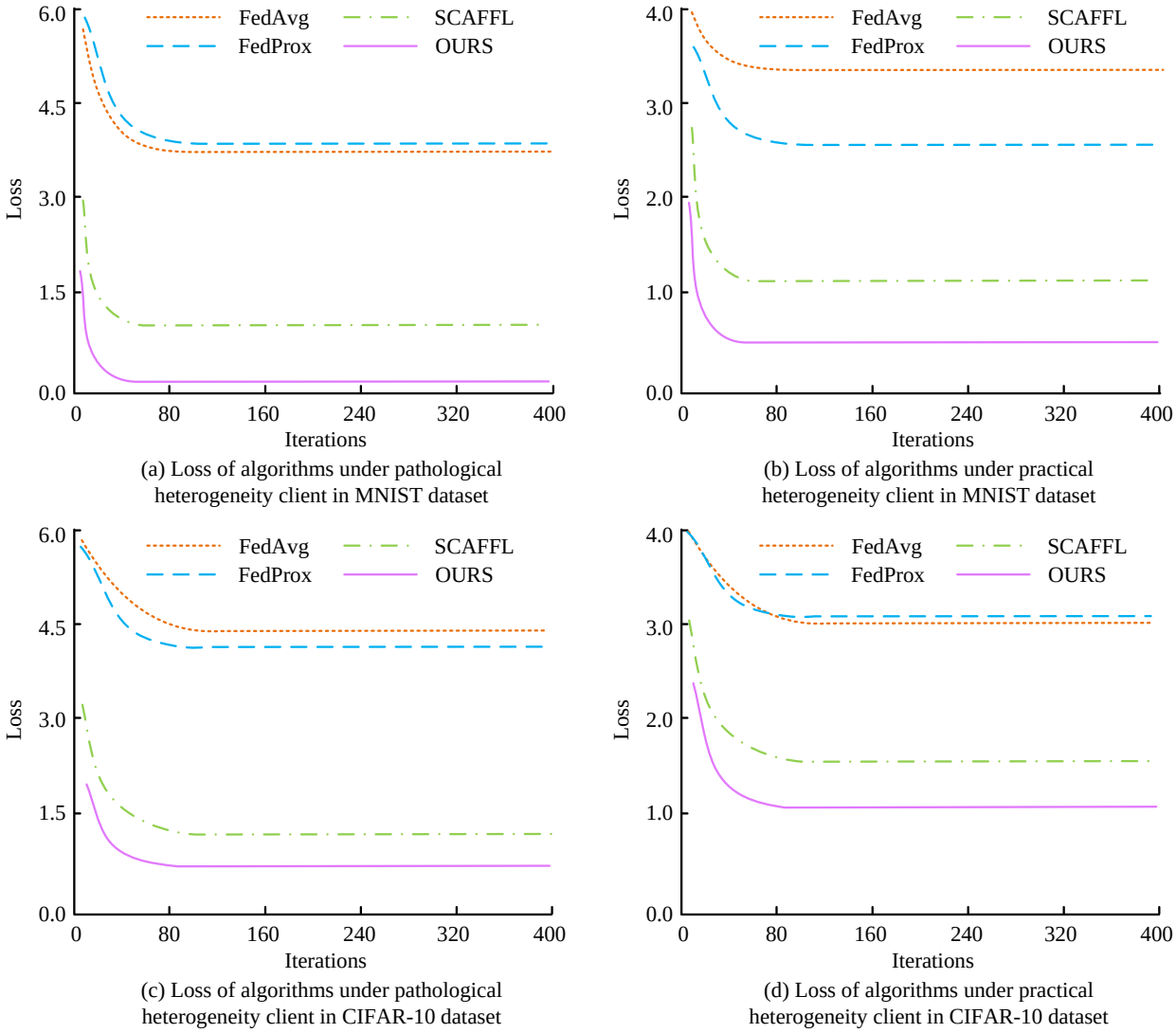
Figure 6: Accuracy of algorithms under two heterogeneous clients in MNIST and CIFAR-10 datasets.

The experimental environment and parameters are denoted in Table 4.

To prove the robustness of the designed algorithm, the accuracy of the algorithm was analyzed on two heterogeneous clients, pathology and real-world scenarios, in the MNIST and CIFAR-10 datasets. Compared with current mainstream algorithms, including FedAvg, FedProx, and Stochastic Controlled Averaging for Federated Learning (ScaFFL), the findings are denoted in Figure 6. In Figure 6 (a), in the MNIST dataset, with an iteration of 160, the accuracy of FedAvg, FedProx, SCAFFL, and the proposed algorithm on the pathological heterogeneity client were 96.24%, 95.16%, 97.68%, and 98.69%, respectively. In Figure 6 (b), when the iteration number was 160, the accuracy of the four algorithms in the MNIST dataset was 96.58%, 97.10%, 97.35%, and

98.14%, respectively, under the actual scenario of heterogeneous clients. In Figure 6 (c), in the CIFAR-10 dataset, with an iteration of 160, the accuracy of FedAvg, FedProx, and SCAFFL under pathological heterogeneity client was 65.37%, 67.69%, and 86.74%, respectively. The accuracy of the proposed algorithm was 90.26%. In Figure 6 (d), under the actual scenario of heterogeneous clients and with 160 iterations, the accuracy of the four algorithms in the CIFAR-10 dataset was 69.38%, 69.12%, 83.56%, and 89.87%, respectively. The findings show that the designed algorithm exhibits higher accuracy and robustness in different data distributions and scenarios.

The loss of different comparison algorithms was analyzed in the MNIST and CIFAR-10 datasets to validate the convergence of the designed algorithm. The findings are denoted in Figure 7. In Figure 7 (a), under the pathological heterogeneity client, when the iteration number was 160, the losses of FedAvg, FedProx, SCAFFL, and the proposed algorithm in the MNIST dataset were 0.348, 0.352, 1.025, and 0.113, respectively. In Figure 7 (b), under the actual scenario of heterogeneous clients, when the iteration number was 160, the losses of the four algorithms were 0.341, 0.326, 1.104, and 0.505, respectively. In Figure 7 (c), in the CIFAR-10 dataset, at an iteration of 160, the losses of the four algorithms under pathological heterogeneity clients were 4.472, 4.210, 1.389, and 0.857, respectively. In Figure 7 (d), under the actual scenario of heterogeneous clients, when the iteration number was 160, the losses of FedAvg, FedProx, and SCAFFL were 3.021, 3.098, and 1.610, respectively, and the loss of the proposed algorithm was 1.024. The findings demonstrate that the designed algorithm can improve the convergence speed of the model and achieve lower losses under different data distributions.



(a) Loss of algorithms under pathological heterogeneity client in MNIST dataset

(b) Loss of algorithms under practical heterogeneity client in MNIST dataset

(c) Loss of algorithms under pathological heterogeneity client in CIFAR-10 dataset

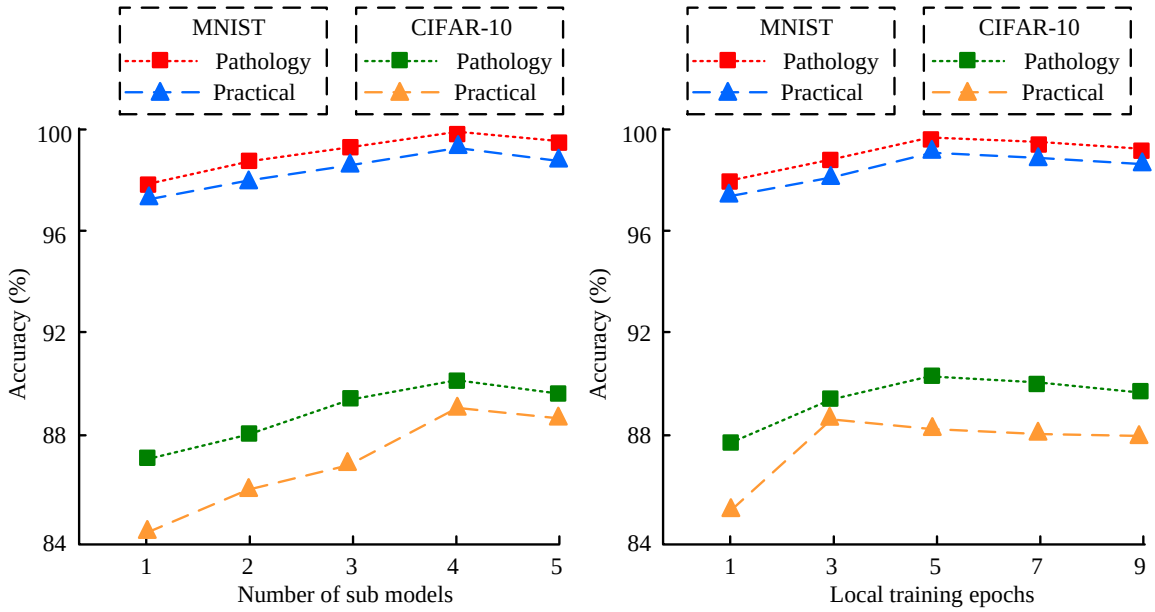
(d) Loss of algorithms under practical heterogeneity client in CIFAR-10 dataset

Figure 7: Loss of algorithms under two heterogeneous clients in MNIST and CIFAR-10 datasets.

Table 5: Comprehensive comparison of model performance and statistical significance.

Dataset	Methods	Pathological Heterogeneity			Real-world Heterogeneity		
		95% CI (Accuracy)	95% CI (Loss)	p-value	95% CI (Accuracy)	95% CI (Loss)	p-value
MNIST	FedAvg	[95.98, 96.44]	[0.335, 0.361]	<0.01	[96.34, 96.76]	[0.328, 0.354]	<0.01
	FedProx	[94.83, 95.45]	[0.339, 0.365]	<0.01	[96.85, 97.31]	[0.313, 0.339]	<0.01
	SCAFFL	[97.47, 97.83]	[1.001, 1.049]	<0.01	[97.13, 97.51]	[1.080, 1.128]	<0.01
	OURS	[98.52, 98.82]	[0.105, 0.121]	/	[98.00, 98.24]	[0.490, 0.520]	/
CIFAR-10	FedAvg	[64.68, 66.00]	[4.350, 4.594]	<0.01	[68.84, 69.86]	[2.909, 3.133]	<0.01
	FedProx	[67.07, 68.23]	[4.098, 4.322]	<0.01	[68.54, 69.64]	[2.986, 3.210]	<0.01
	SCAFFL	[86.31, 87.11]	[1.367, 1.411]	<0.01	[83.09, 83.97]	[1.588, 1.632]	<0.01
	OURS	[89.97, 90.51]	[0.841, 0.873]	/	[89.57, 90.13]	[1.010, 1.038]	/

Note: p<0.01 indicates reaching a highly significant level.



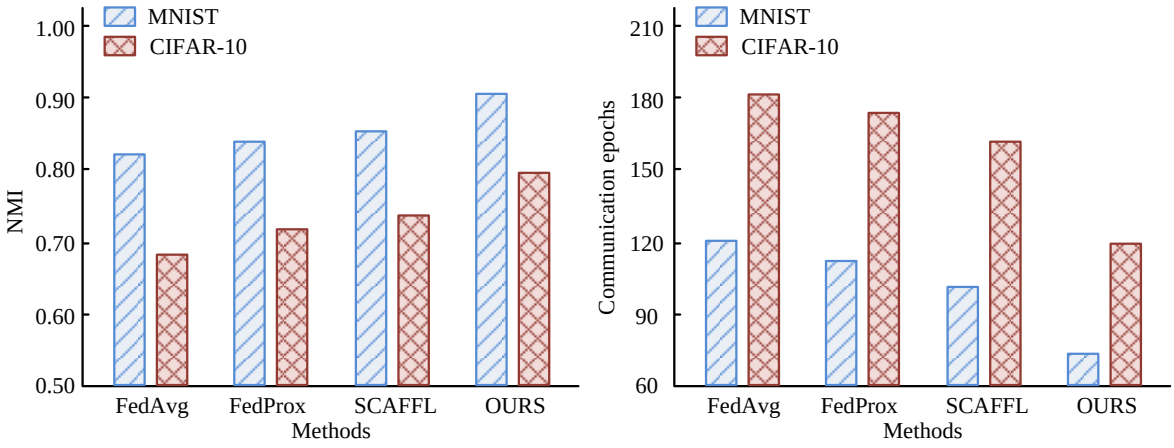
(a) The impact of different numbers of sub models on the accuracy of the algorithm

(b) The impact of different local training epochs on the accuracy of the algorithm

Figure 8: The impact of different numbers of sub models and local training epochs on the accuracy of improved algorithms.

This study conducted a statistical analysis of the accuracy loss of the proposed algorithm under two heterogeneous clients, pathology and real-world scenarios, in the MNIST and CIFAR-10 datasets. The results are shown in Table 5. The statistical results show that the method proposed in this study significantly outperforms the baseline algorithm in both datasets and heterogeneous scenarios. On the MNIST dataset, the accuracy and 95% confidence interval of the proposed method were [98.52, 98.82] and [98.00, 98.24], respectively, demonstrating applicability to heterogeneous pathology and real-world scenarios, and the 95% confidence intervals for loss values were [0.105, 0.121] and [0.490, 0.520], respectively. On the more complex CIFAR-10 dataset, the proposed method also outperformed all baselines in terms of confidence intervals. All p-values compared were less than 0.001, indicating that the performance improvement is highly statistically significant.

A comparative analysis was conducted on the accuracy of the MNIST and CIFAR-10 datasets under different numbers of sub-models and local training epochs, to investigate the impact of these factors on the performance improvement of the proposed algorithm. The findings are denoted in Figure 8. In Figure 8 (a), under the pathological heterogeneity client, when the number of sub-models was 4, the accuracy of the designed algorithm in the two datasets was 99.82% and 90.33%, respectively. In actual heterogeneous client scenarios, their accuracy rates were 99.47% and 89.11%, respectively. In Figure 8 (b), when the local training epochs were 5, the accuracy of the proposed algorithm in two datasets was 99.86% and 90.62% respectively under pathological heterogeneity client. In actual scenarios with heterogeneous clients, the accuracy rates were 99.43% and 88.94%, respectively. Research has found that increasing the number of sub-models and local training rounds can improve algorithm performance.



(a) Comparison of NMI using different methods on two datasets

(b) Comparison of communication epochs using different methods on two datasets

Figure 9: Integration quality and clustering efficiency of different methods on two datasets.

3.2 Performance evaluation of intelligence dynamic data integration and clustering methods

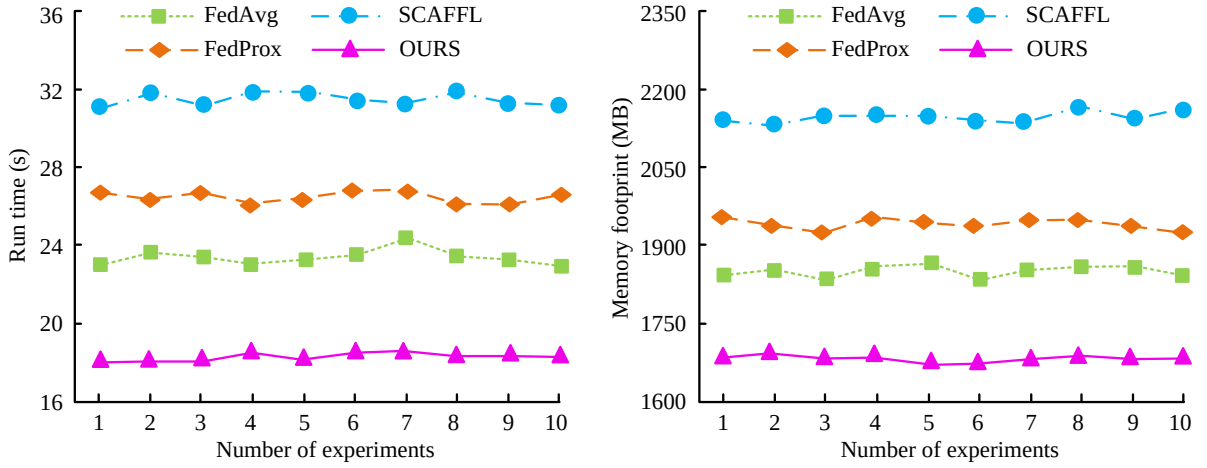
The ensemble quality and clustering efficiency of the proposed method were analyzed in the MNIST and CIFAR-10 datasets, and compared with FedAvg, FedProx, and SCAFFL. The indicator for evaluating integration quality was Normalized Mutual Information (NMI), and the clustering efficiency indicator was communication epochs. The findings are denoted in Figure 9. In Figure 9 (a), in the MNIST dataset, the NMIs of FedAvg, FedProx, SCAFFL, and the designed algorithm were 0.82, 0.84, 0.86, and 0.91, respectively. The NMIs in the CIFAR-10 dataset were 0.68, 0.71, 0.73, and 0.79, respectively. In Figure 9 (b), the communication epochs of FedAvg, FedProx, and SCAFFL in the MNIST dataset were 121, 112, and 101, respectively, and in the CIFAR-10 dataset were 179, 171, and 162, respectively. Compared with it, the proposed algorithm had 75 and 119 communication epochs in the two datasets, respectively. The findings show that the designed algorithm can ensure high integration quality while reducing communication overhead, verifying its efficiency in dynamic data integration and clustering tasks.

Ablation experiments were conducted to verify the contribution of each core module to the overall performance of the dynamic data integration and clustering methods. The findings are presented in Table 6. In ablation analysis, the roles of each module are as follows: the decomposition and combination mechanism adapts to heterogeneous data distributions by dividing sub models; The pre-training mechanism utilizes historical features to optimize model initialization and accelerate convergence; Hierarchical clustering identifies intrinsic relationships between clients through a multi-level structure, improving grouping stability and aggregation quality, which is superior to methods that rely solely on planar partitioning; The dynamic weighting mechanism adjusts client contributions based on data timeliness to alleviate concept drift. In Table 6, on the MNIST dataset, the F1 score of the complete method was 0.924 and the precision was 0.931, significantly higher than other configurations. On the CIFAR-10 dataset, the F1 score and precision of the complete method were 0.802 and 0.810, respectively. The F1 score and precision of the baseline method were the lowest, with values of 0.852 and 0.806 in the MNIST dataset and 0.723 and 0.730 in the CIFAR-10 dataset, respectively. The findings denote that each core module contributes significantly to the improvement of model performance.

The study analyzed the running time and memory usage of the designed method in a practical scenario of a medical IoT, and compared it with other methods. The findings are denoted in Figure 10. In Figure 10 (a), the average running times of FedAvg, FedProx, SCAFFL, and the proposed algorithm were 23.42s, 26.75s, 31.52s, and 18.23s, respectively. Compared with the comparative algorithm, the running time of the proposed algorithm was reduced by 22.16%, 31.85%, and 42.16%, respectively. In Figure 10 (b), the average memory usage of FedAvg, FedProx, and SCAFFL was 1853MB, 1926MB, and 2157MB, respectively. Compared with them, the proposed algorithm had an average memory usage of 1681MB, which was reduced by 9.28%, 12.72%, and 22.07%, respectively. The outcomes demonstrat that the designed algorithm can substantially decrease the consumption of computational resources while maintaining performance, thereby verifying its practicality and deployment benefits in IoT scenarios with limited resources.

Table 6: Results of ablation experiment.

Method configuration	MNIST dataset				CIFAR-10 dataset			
	Precision	Recall	F1 score	(*p*-value vs. FedAvg)	Precision	Recall	F1 score	(p-value vs. FedAvg)
Complete method	0.931	0.918	0.924	<0.001	0.810	0.795	0.802	<0.001
No decomposition combination	0.879	0.863	0.871	<0.001	0.742	0.728	0.735	<0.001
No pre training	0.900	0.886	0.893	<0.001	0.775	0.761	0.768	<0.001
Non hierarchical clustering	0.893	0.879	0.886	<0.001	0.760	0.745	0.752	<0.001
No dynamic weight	0.908	0.894	0.901	<0.001	0.786	0.772	0.779	<0.001
Baseline method (FedAvg)	0.806	0.845	0.852	/	0.730	0.716	0.723	/

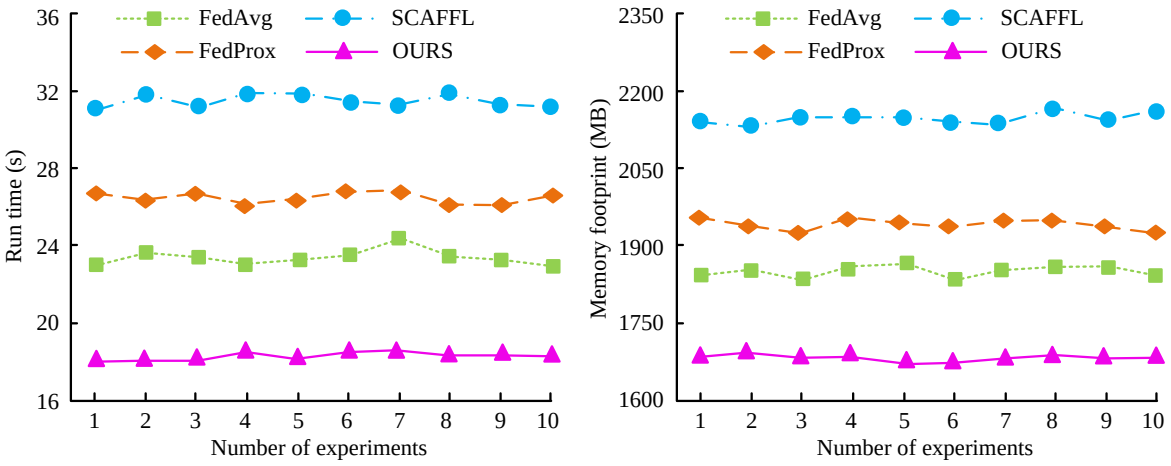


(a) Comparison of run time of different methods in practical scenarios

(b) Comparison of memory usage of different methods in practical scenarios

Figure 10: Comparison of runtime and memory usage of different methods in practical scenarios.

To verify the effectiveness of the proposed method in practical datasets, an analysis was conducted on the running time and average memory of different methods under different numbers of clients in the PhysioNet dataset. The PhysioNet dataset is a widely recognized real-world clinical time series dataset containing records of 12000 ICU patients. The results are shown in Figure 11. In Figure 11 (a), when the number of clients was 20, the running times of FedAvg, FedProx, and SCAFFL were 28.45s, 32.11s, and 38.94s, respectively, and the running time of the proposed method was 21.08s. When the number of clients increased to 100, the running times of the four methods were 62.34s, 71.89s, 88.56s, and 45.12s, respectively. In Figure 11 (b), when the number of clients was 20, the average memory of FedAvg, FedProx, SCAFFL, and the proposed method was 2105MB, 2189MB, 2455MB, and 1950MB, respectively. When the number of clients reached 100, the average memory was 3521MB, 3744MB, and 1950MB, respectively. B, 4455MB, and 2850MB. The results show that the proposed method effectively reduces computational and storage costs on real medical datasets, demonstrating superior scalability and practicality.



(a) Comparison of run time of different methods in practical scenarios

(b) Comparison of memory usage of different methods in practical scenarios

Figure 11: Run time and average memory of different methods under different numbers of clients.

Table 7: Model complexity comparison across different datasets.

Methods	MNIST	CIFAR-10	PhysioNet	MNIST	CIFAR-10	PhysioNet
	Number of Parameters (Millions)			FLOPs (MegaFLOPs)		
FedAvg	4.21	23.45	1.85	8.45	125.67	3.72
FedProx	4.21	23.47	1.85	8.47	125.72	3.73
SCAFFL	4.26	23.51	1.87	8.52	126.05	3.76
OURS	5.18	25.83	2.21	10.31	135.42	4.45

Further analysis was conducted on the parameter count and Floating-Point Operations (FLOPs) of different methods on the MNIST, CIFAR-10, and PhysioNet datasets. The results are shown in Table 7. In Table 7, on the MNIST dataset, the proposed method had a parameter size of 5.18M and FLOPs of 10.31M, which were approximately 23% and 22% higher than FedAvg, respectively. The complexity of the proposed method on the real clinical dataset PhysioNet also maintained a similar increase. This controllable increase in complexity, compared to the specific performance improvements obtained in the previous experiments, demonstrates that the proposed method achieves a good balance between efficiency and performance.

4 Discussion

In this study, a dynamic integration and clustering method for intelligence data based on an improved FL algorithm was proposed. The experimental results demonstrated significant improvements in accuracy, convergence speed, and communication efficiency compared to several existing approaches.

In terms of accuracy, this method achieved accuracies of 98.69% and 90.26% respectively on the MNIST and CIFAR-10 datasets for pathological heterogeneous clients, and 98.14% and 89.87% respectively in real-world heterogeneous clients. Compared with the real-time medical data processing method proposed in reference [6], the accuracy of this method was improved by about 6-8 percentage points. This improvement is mainly due to the dynamic sub model aggregation mechanism, which decomposes the global model into multiple specialized sub models, enabling the model to better adapt to the data distribution characteristics of different clients.

In terms of convergence performance, this method only required 75 communication rounds to converge on the MNIST dataset and 119 communication rounds on the CIFAR-10 dataset. In contrast, the privacy preserving FL method in reference [10] required 121 and 179 communication epochs respectively in similar tasks. The improvement in convergence speed is mainly due to the introduction of pre-training mechanisms, which utilize historical data feature information for model initialization, enabling the model to have a good parameter foundation in the early stages of training, thereby accelerating the training process.

In terms of communication efficiency, this method had a running time of 18.23 seconds and a memory usage of 1681MB in actual medical IoT scenario testing. Compared with the method in reference [12], the running time was reduced by 27% and the memory usage was reduced by 12%. This improvement is due to the application of hierarchical similarity clustering technology, which significantly reduces unnecessary communication overhead by intelligently grouping clients.

However, this method also has some limitations. Firstly, due to the adoption of a multi submodel architecture and clustering process, its computational complexity is relatively high, which may limit its application in resource constrained environments. Secondly, the method is sensitive to hyperparameter settings, especially the selection of the number of clusters K and learning rate, which can significantly affect performance. Compared with the heterogeneous IoT FL framework in reference [13], this method improved

accuracy by about 4% on the CIFAR-10 dataset, but increased computational load by about 15%. This indicates that while pursuing performance improvement, a balance needs to be struck between accuracy and computational efficiency. Compared with the TEE based method in reference [8], although this method avoided the dependence on dedicated hardware, it may be slightly inadequate in combating advanced security threats.

In summary, the proposed method has achieved significant improvements in accuracy, convergence speed, and communication efficiency through the organic combination of dynamic submodel aggregation, pre training mechanism, and hierarchical clustering. Future research will focus on developing adaptive hyperparameter optimization strategies, reducing computational complexity, and exploring decentralized aggregation mechanisms to enhance system robustness.

5 Conclusion

The study proposes a dynamic integration and clustering method for intelligence data based on an improved FL algorithm to strengthen the effectiveness of the FL algorithm in intelligence data processing. The findings denote that the designed method can effectively solve the challenges of slow convergence speed, poor adaptability to dynamic data, and low clustering efficiency in FL. It not only improved precision and convergence speed, but also reduced computational resource consumption, making it suitable for practical application scenarios such as medical IoT.

6 Funding

The research is supported by: the Science and Technology Project of State Grid Shanxi Electric Power Company “Research and Application of Key Technologies for Information Extraction and Association Based on Federated Learning Large Models” (Grant number: 5205M024000K).

References

[1] Iqbal H. Sarker. Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. Annals of Data Science, 10(6):1473-1498, 2023. https://doi.org/10.1007/s40745-022-00444-2

[2] Zhijuan Zong, and Yu Guan. AI-driven intelligent data analytics and predictive analysis in Industry 4.0: Transforming knowledge, innovation, and efficiency. Journal of the Knowledge Economy, 16(1):864-903, 2025. https://doi.org/10.1007/s13132-024-02001-z

[3] Lei Ren, Yingjie Li, Xiaokang Wang, Jin Cui, and Lin Zhang. An ABGE-aided manufacturing knowledge graph construction approach for heterogeneous IIoT data integration. International Journal of Production Research, 61(12):4102-4116, 2023. https://doi.org/10.1080/00207543.2022.2042416

[4] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. International Journal of Machine Learning and Cybernetics, 14(2):513-535, 2023. https://doi.org/10.1007/s13042-022-01647-y

[5] Qiang Yang, Anbu Huang, Lixin Fan, Chee Seng Chan, Jian Han Lim, Kam Woh Ng, Ding Sheng Ong, and Bowen Li. Federated learning with privacy-preserving and model IP-right-protection. Machine Intelligence Research, 20(1):19-37, 2023. https://doi.org/10.1007/s11633-022-1343-2

[6] Kehua Guo, Tianyu Chen, Sheng Ren, Nan Li, Min Hu, and Jian Kang. Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 21(4):869-879, 2022. https://doi.org/10.1109/TCBB.2022.3185395

[7] Tomer Gafni, Nir Shlezinger, Kobi Cohen, Yonina C. Eldar, and H. Vincent Poor. Federated learning: A signal processing perspective. IEEE Signal Processing Magazine, 39(3):14-41, 2022. https://doi.org/10.1109/MSP.2021.3125282

[8] Abbas Yazdinejad, Ali Dehghantanha, and Gautam Srivastava. AP2FL: Auditable privacy-preserving federated learning framework for electronics in healthcare. IEEE Transactions on Consumer Electronics, 70(1):2527-2535, 2023. https://doi.org/10.1109/TCE.2023.3318509

[9] Guanming Bao, and Ping Guo. Federated learning in cloud-edge collaborative architecture: Key technologies, applications and challenges. Journal of Cloud Computing, 11(1):94-115, 2022. https://doi.org/10.1186/s13677-022-00377-4

[10] Ruijin Wang, Jinshan Lai, Zhiyang Zhang, Xiong Li, Pandi Vijayakumar, and Marimuthu Karuppiah. Privacy-preserving federated learning for internet of medical things under edge computing. IEEE Journal of Biomedical and Health Informatics, 27(2):854-865, 2022. https://doi.org/10.1109/JBHI.2022.3157725

[11] Pushpita Chatterjee, Debashis Das, and Danda B. Rawat. Federated learning empowered recommendation model for financial consumer services. IEEE Transactions on Consumer Electronics, 70(1):2508-2516, 2023. https://doi.org/10.1109/TCE.2023.3339702

[12] Mahmuda Akter, Nour Moustafa, Timothy Lynar, and Imran Razzak. Edge intelligence: Federated learning-based privacy protection framework for smart healthcare systems. IEEE Journal of Biomedical and Health Informatics, 26(12):5805-5816, 2022. https://doi.org/10.1109/JBHI.2022.3192648

[13] Demin Gao, Haoyu Wang, Xiuzhen Guo, Lei Wang, Guan Gui, and Weizheng Wang. Federated learning based on CTC for heterogeneous internet of things. IEEE Internet of Things Journal, 10(24):22673-22685, 2023. https://doi.org/10.1109/JIOT.2023.3305189

[14] Zhiguo Qu, Lailei Zhang, and Prayag Tiwari. Quantum fuzzy federated learning for privacy protection in intelligent information processing. IEEE Transactions on Fuzzy Systems, 33(1):278-289, 2024. https://doi.org/10.1109/TFUZZ.2024.3419559

[15] Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. ACM Transactions on Intelligent Systems and Technology (TIST), 13(4):1-23, 2022. https://doi.org/10.1145/3501813

[16] Shaoxiong Ji, Yue Tan, Teemu Saravirta, Zhiqin Yang, Yixin Liu, Lauri Vasankari, Shirui Pan, Guodong Long, Anwar Walid, Vasankari L, and Walid A. Emerging trends in federated learning: From model fusion to federated x learning. International Journal of Machine Learning and Cybernetics, 15(9):3769-3790, 2024. https://doi.org/10.1007/s13042-024-02119-1

[17] Ali Hatamizadeh, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, Mona G Flores, Jan Kautz, Daguang Xu, and Holger R. Roth. Do gradient inversion attacks make federated learning unsafe. IEEE Transactions on Medical Imaging, 42(7):2044-2056, 2023. https://doi.org/10.1109/TMI.2023.3239391

[18] Weimin He, and Lei Zhao. Application of federated learning algorithm based on K-means in electric power data. Journal of New Media, 4(4):191-203, 2022. https://doi.org/10.32604/jnm.2022.032994

[19] Chaoli Sun, Xiaojun Wang, Junwei Ma, and Gang Xie. A composition-decomposition based federated learning. Complex & Intelligent Systems, 10(1):1027-1042, 2024. https://doi.org/10.1007/s40747-023-01198-x

[20] Mansoor Ali, Faisal Naeem, Muhammad Tariq, and Georges Kaddoum. Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. IEEE Journal of Biomedical and Health Informatics, 27(2):778-789, 2022. https://doi.org/10.1109/JBHI.2022.3181823

[21] Pradyumna Kumar Tripathy, Anurag Shrivastava, Varsha Agarwal, Devangkumar Umakant Shah, Chandra Sekhar Reddy L., and S.V. Akilandeeswari. Federated learning algorithm based on matrix mapping for data privacy over edge computing. International Journal of Pervasive Computing and Communications, 20(5):633-647, 2024. https://doi.org/10.1108/IJPCC-03-2022-0113

[22] Zhiyuan Wang, Hongli Xu, Jianchun Liu, Yang Xu, He Huang, and Yangming Zhao. Accelerating federated learning with cluster construction and hierarchical aggregation. IEEE Transactions on Mobile Computing, 22(7):3805-3822, 2022. https://doi.org/10.1109/TMC.2022.3147792

[23] Jose A. Carrillo, Nicolas Garcia Trillos, Sixu Li, and Yuhua Zhu. FedCBO: Reaching group consensus in clustered federated learning through consensus-based optimization. Journal of Machine Learning Research, 25(214):1-51, 2024. DOI: 10.48550/arXiv.2305.02894

[24] Yongheng Deng, Feng Lyu, Tengxi Xia, Yuezhi Zhou, Yaoxue Zhang, Ju Ren, and Yuanyuan Yang. A communication-efficient hierarchical federated learning framework via shaping data distribution at edge. IEEE/ACM Transactions on Networking, 32(3):2600-2615, 2024. https://doi.org/10.1109/TNET.2024.3363916

[25] Chengtian Ouyang, Yehong Li, Jihong Mao, Donglin Zhu, Changjun Zhou, and Zhenyu Xu. Enhancing federated learning with dynamic weight adjustment based on particle swarm optimization. Discover Computing, 27(1):35-52, 2024. https://doi.org/10.1007/s10791-024-09478-x

Kernel-PCA + RBM Feature Extraction with Optimised Naïve Bayes for Intrusion Detection in Impaired Wireless Sensor Networks

Shengguo Guo*, Dandan Xing
School of Information Engineering, Zhenzhou College of Finance and Economics, Zhengzhou Henan 450000, China
E-mail: guosg@imte.ac.cn

Keywords: network impairment, wireless sensor network, data dimensionality reduction, feature extraction, abnormal intrusion detection, naive bayes

Received: August 14, 2025

The high-dimensional communication data generated by wireless sensor networks often contains substantial redundant and irrelevant information, which hampers the effective retention of critical features. Consequently, the characteristics of network impairment states and abnormal intrusion behaviors become intertwined and difficult to distinguish, ultimately compromising the accuracy of intrusion detection. Therefore, this paper studies the method of abnormal intrusion detection of wireless sensor network communication under network impairment. First, global node perception is achieved through the wireless sensor network networking model to obtain high-dimensional communication data. Second, the kernel principal component analysis (KPCA) method is used to perform nonlinear dimensionality reduction on the data, significantly reducing the data dimension and computational complexity while retaining the key information in the data. Subsequently, a restricted Boltzmann machine (RBM) is introduced to extract the deep features of the dimensionality-reduced data to distinguish the feature differences between network impairment states and abnormal intrusions. Finally, a high-precision abnormal intrusion detection is achieved through an optimized naive Bayes classifier. This classifier effectively improves the anti-interference ability under network impairment states by feature weighting and micro conditional probability optimization, highlights key features, and realizes abnormal intrusion detection. The experiment was conducted on a WSN dataset containing 50000 records, simulating a damaged scenario with a 30% packet loss rate and a 40% bandwidth limitation. The results showed that the proposed method reduced the data dimensionality from 90 to 15 dimensions, with a variance retention rate of 94.7%; In the detection of 10 types of attacks, the F1 value reaches 0.92, which is better than CNN (0.60) and association rules (0.62); At a 75% network damage rate, the false positive rate is only 5%, with accuracy and recall rates of 0.94 and 0.86, respectively, and a single sample prediction time of only 0.21 ms. This method maintains high detection accuracy while having low computational overhead and strong robustness, making it suitable for WSN security protection in complex damaged environments.

Povzetek: Prispevek predstavlja učinkovito metodo zaznavanja vdorov v brezžičnih senzorskih omrežjih z visoko natančnostjo in nizko računsko zahtevnostjo tudi ob okvarah omrežja.

1 Introduction

Wireless sensor networks (WSNs) have been widely applied in key fields such as industrial monitoring, environmental perception, and intelligent transportation due to their advantages [1] of self-organization, low cost, and high flexibility [2]. Abnormal intrusion detection in wireless sensor network communication [3] aims to identify and defend against malicious behaviors such as illegal node access, data tampering, and denial-of-service attacks in real time [4], ensuring the security and reliability of network data transmission [5]. However, in actual application scenarios, the network often faces multiple impairment risks such as channel interference, node failure, and resource constraints [6]. Conducting research on abnormal intrusion detection under network impairment not only concerns the stable operation of the sensor network itself but is also crucial for avoiding serious consequences such as industrial accidents and data leakage [7, 8], having important practical significance for building a solid security defense line for the Internet of Things and promoting the healthy development of the intelligent industry [9].

Many scholars have conducted research on intrusion detection for abnormal communications in wireless sensor networks. For example, Niccolò et al. reconstructed the data packets transmitted by wireless

sensor nodes through an autoencoder, identified abnormal traffic based on the reconstruction error, and achieved the detection of unknown intrusions [10]. However, the performance of autoencoders heavily depends on the reconstruction mode of normal traffic. Under network damage, abnormal traffic generated by channel interference or node failure may have reconstruction errors that are highly similar to malicious intrusion behavior, making it difficult for the model to effectively distinguish at the feature representation level, resulting in poor detection performance. Karrothu et al. adopted an end-cloud-fog detection structure. They collected sensor data at the endpoint layer and transmitted it. The data of the wireless sensor network was transformed at the cloud computing layer through the Yeo-Johnson transform, and feature selection was carried out with the Kulczynski similarity. Then, the selected features were sent to an ensemble classifier optimized by the GDO (Gazelle-Dog) algorithm, and intrusion detection was completed using the ensemble classifier at the fog computing layer [11]. However, the Kulczynski similarity relies on the statistical relationship between data for feature screening. When a new attack pattern or environmental mutation occurs in the wireless sensor network, the relationship between data features changes, and the existing similarity calculation may misjudge the importance of features, thus missing key abnormal features and reducing the detection accuracy. Arkan et al. combined the wireless sensor network with the software-defined network by constructing a software-defined wireless sensor network architecture (SDWSN). The sensor runs an unsupervised intrusion detection algorithm module locally, clusters and analyzes the data based on entropy and cumulative point similarity, and sends the results to the SDWSN controller. The controller comprehensively analyzes the data analysis results of each region to determine whether the data is abnormal and complete abnormal intrusion detection [12]. However, in the case of network damage, this method uses entropy and cumulative point similarity as clustering criteria. In the case of network damage, packet loss and noise interference can significantly affect the stability of entropy and cumulative point similarity calculations, causing deviations in these clustering criteria and ultimately incorrectly dividing normal and abnormal data, affecting the accuracy of intrusion detection. Mutambik adopted the IoT-FIDS (Internet of Things Intrusion Detection Based on Data Streams) lightweight framework to achieve abnormal intrusion detection. This framework captures details such as node communication patterns and service usage by checking data streams, and only analyzes benign traffic during the detection process to identify abnormal behaviors, avoiding the dependence on pre-labeled data and a large amount of computing power. While reducing resource consumption, it can accurately detect most abnormal traffic, reduce false alarms, and provide practical protection for network security [13]. However, this method cannot highlight the role of key features and is difficult to accurately distinguish between normal traffic changes and malicious behaviors, resulting in misjudgment.

In addition to the research on detection architecture mentioned above, from the perspective of data processing and feature engineering, existing works have also attempted to introduce various technologies. In terms of data dimensionality reduction, Zhang B et al. [14] used t-SNE for nonlinear dimensionality reduction, which focuses on preserving the local neighborhood structure of data points and mapping high-dimensional features (such as texture features extracted by GLCM) to a low dimensional space. However, such methods typically focus on maintaining the topological structure of data for visualization, with high computational complexity and a lack of clear inverse mapping for dimensionality reduction results, making it difficult to directly serve efficient online detection tasks. Shen Z et al. [15] used UMAP technology to reduce and visualize high-dimensional features, effectively evaluating the class separability of features in low dimensional space. However, the dimensionality reduction results of UMAP have randomness, and their output is sensitive to initialization parameters, which can lead to inconsistent feature representations after dimensionality reduction at different times or network states, seriously damaging the stability of online detection models. In terms of deep feature extraction, Alshehri et al. [16] proposed a model that combines Wasserstein GAN and autoencoder (WGAN-AE), which utilizes autoencoder reconstruction to learn robust latent feature representations. However, the training process of such generative models is unstable, and under the noise interference introduced by network damage, the dynamic balance between the generator and discriminator is more difficult to maintain, resulting in a decrease in the reliability of feature extraction. Brian W et al. [17] constructed a hybrid model integrating Transformer and random forest, which utilizes the self attention mechanism of Transformer to dynamically screen and weight key features. However, this model has high computational complexity and strict hardware resource requirements, making it difficult to deploy on wireless sensor network nodes with limited computing power; At the same time, its powerful feature filtering ability highly relies on a large amount of high-quality annotated data. In practical scenarios where the network is damaged, attack patterns are variable, and labels are scarce, its performance faces severe challenges.

The summary of research on anomaly intrusion detection in wireless sensor networks is shown in Table 1.

Table 1: Summary of related research on wireless sensor network anomaly intrusion detection

Author(s) (Year)	Research Method	Scenario	Key Results	Limitations
Niccolò et al. (2024)	Autoencoder reconstructs data packets transmitted by wireless sensor nodes, identifies anomalies based on reconstruction error	Wireless sensor networks	Detects unknown intrusions	Performance heavily depends on normal traffic patterns; struggles to distinguish between channel interference and malicious intrusion under network impairment.
Karrothu et al. (2025)	End-cloud-fog architecture, Yeo-Johnson transform + Kulczynski similarity + GDO-optimized ensemble classifier	Wireless sensor networks	Achieves distributed intrusion detection with optimized resources	Feature selection relies on static statistical relationships; poor adaptability to new attack patterns or environmental changes.
Arkan et al. (2023)	SDWSN architecture, clustering analysis based on entropy and cumulative point similarity	Software-defined wireless sensor networks	Achieves unsupervised, hierarchical intrusion detection	Clustering criteria are sensitive to network impairment; packet loss and noise affect computational stability.
Mutambik (2024)	IoT-FIDS lightweight framework, data flow pattern inspection	IoT data streams	Low resource consumption, reduced false positives	Unable to highlight key features; normal fluctuations in complex environments are easily confused with attacks.
Zhang B et al. (2025)	t-SNE nonlinear dimensionality reduction	Defect detection (texture features)	Effectively preserves local data structure for visualization	High computational complexity, lacks inverse mapping, difficult to use for online detection.
Shen Z et al. (2024)	UMAP dimensionality reduction and visualization	Optical performance monitoring data	Effectively evaluates feature class separability	Dimensionality reduction results are stochastic, compromising the stability of online detection models.
Alshehri et al. (2025)	WGAN-AE hybrid model	IoT intrusion detection data	Learns robust latent feature representations	Unstable training process, sensitive to noise interference.
Brian W et al. (2024)	Transformer and random forest hybrid model	Satellite ground station networks	Dynamically selects and weights key features	High computational complexity, significant resource demands, relies on large amounts of labeled data.

Based on the systematic analysis of existing research (Table 1), it can be found that the current wireless sensor network anomaly intrusion detection faces three core problems: firstly, in the state of network damage, existing methods are difficult to distinguish the anomalous features of physical damage and malicious intrusion, resulting in high false alarm rates; Secondly, there is a lack of efficient and stable nonlinear feature processing solutions. Existing dimensionality reduction and feature extraction methods are either computationally complex, have unstable results, or have high resource requirements, which cannot meet the real-time online detection needs; The third issue is that the classifier lacks anti-interference ability in damaged environments, making it unable to adaptively highlight key features, resulting in a decrease in detection accuracy. Therefore, in order to systematically solve the problems faced by wireless sensor network communication anomaly intrusion detection under network damage and improve the accuracy of wireless sensor network communication anomaly intrusion detection, this paper studies a method for detecting wireless sensor network communication anomaly intrusion under network damage. It should be noted that the "network damage state" focused in this article specifically refers to systematic communication quality collapse scenarios caused by physical damage, malicious preemption, or sudden congestion with high random packet loss rates (>30%), malicious bandwidth limitations, and critical node failures. Unlike traditional static background constraints such as limited resources or environmental interference, this state leads to highly distorted and incomplete data transmission in the network, causing deep confusion between intrusion features and physical damage features, thereby undermining the assumption that existing detection methods rely on "basic data reliability". The specific technical route is as follows:

(1) Global perceptible data acquisition: By constructing a wireless sensor network networking model and based on the probability formula for node automatic identification, it ensures that the status monitoring and communication data collection of all nodes in the entire domain can still be achieved even when the network is partially damaged, providing a reliable data foundation for subsequent analysis.

(2) KPCA Nonlinear Data Dimensionality Reduction: To address the difficulty of transmitting and processing high-dimensional data in damaged networks, the Kernel Principal Component Analysis (KPCA) method is adopted. Nonlinear mapping of raw data to low dimensional space using Gaussian radial basis kernel function significantly reduces data dimensionality and computational complexity while preserving key structural information, alleviating resource pressure.

(3) RBM deep feature extraction: To effectively distinguish the mixed features of network damage and malicious intrusion, a restricted Boltzmann machine (RBM) is used to perform unsupervised deep feature learning on the reduced dimensional data. By using its energy model probability distribution, abstract features that can characterize the essence of abnormal intrusion are extracted to enhance the discriminative ability of the features.

(4) Optimizing Naive Bayes classification decisions: In response to noise interference in damaged environments, perform dual optimization on the Naive Bayes classifier: first, introduce a feature weighting mechanism to highlight the contribution of key features; Secondly, implement micro conditional probability optimization to improve the robustness of the model by iteratively modifying the conditional probability estimation. Ultimately, the optimized classifier is utilized to achieve high-precision and low false positive anomaly intrusion detection.

In summary, this article constructs an end-to-end solution through a progressive technical route of "KPCA dimensionality reduction → RBM feature extraction → optimized naive Bayes classification". This solution systematically addresses the full chain challenges from data preprocessing, feature engineering to final decision-making, in order to improve the accuracy, robustness, and real-time performance of wireless sensor network communication anomaly intrusion detection under network damage conditions.

2 Data dimensionality reduction and deep feature extraction of wireless sensor network communication network under damaged state

In wireless sensor networks, in addition to network damage caused by communication anomaly intrusion, physical interference, environmental interference, and energy depletion can all lead to a damaged state of the wireless sensor network. In order to still achieve accurate detection of communication anomaly intrusion in the wireless sensor network under the damaged state, it is necessary to ensure that the entire wireless sensor network is in a perceivable state, that is, each node in the network can be effectively monitored and detected [18]. Through the design of the networking model, ensure the coverage area and communication quality of the sensor network, so as to provide a reliable data source for the subsequent automatic identification of abnormal nodes [19]. In this context, this study focuses on three core objectives: firstly, to verify whether the nonlinear dimensionality reduction method based on kernel principal component analysis can significantly reduce data dimensionality and computational complexity while preserving key information, thereby improving detection real-time performance; The second is to explore the ability of restricted Boltzmann machines to extract deep features from reduced dimensional data to distinguish between network damage and malicious intrusion behavior; The third is to evaluate the performance of the

Naive Bayes classifier with feature weighting and micro conditional probability optimization in maintaining high detection accuracy and low false alarm rate in network damaged environments, and compare its advantages and disadvantages with traditional and deep learning methods. Therefore, this article constructs an end-to-end detection framework that integrates global perception data acquisition, KPCA dimensionality reduction, RBM feature extraction, and optimized Naive Bayes classification, aiming to improve the accuracy, robustness, and real-time performance of anomaly intrusion detection in complex damage scenarios.

2.1 Data acquisition of wireless sensor network under damaged state

To ensure accurate anomaly intrusion detection even in the event of network damage, it is first necessary to ensure that the entire wireless sensor network (WSN) is in a globally perceptible state, where every node in the network can be effectively monitored by the monitoring system. Therefore, this article constructs a geometric perception-based node identifiability model to evaluate the monitoring coverage of each node in the event of network damage. For the convenience of theoretical modeling and analysis, this study simplifies the network deployment area into a rectangular region. Let the entire wireless sensor network be within the rectangle $M \times M$. If the sensing coordinate of a certain wireless sensor network node i within $M \times M$ is (x_i, y_i), whether it can be recognized when abnormal communication occurs in this wireless sensor network node depends on whether it is located within the optimal perception radius of the central node J, regardless of whether it has a logical association with the central node. Let (x, y) represent the coordinates of the automatic identification node. The formula for the condition for the abnormal node i of the wireless sensor network communication to be automatically identified is:

$$P(i, j) = 1, |i, j| < r \tag{1}$$

In the formula, $P(i, j)$ is the automatic identification probability, where a value of 1 indicates that it can be recognized and 0 indicates that it cannot be recognized, $|i, j|$ is the physical distance between node i and j, reflecting the spatial relationship between nodes, and r is the optimal sensing radius of the central node j, which is determined by the node's communication capability, channel quality, and environmental interference. It can usually be estimated through actual measurements or link budget models, representing the maximum effective range that the node can reliably perceive and communicate under specific network and environmental configurations.

The expression of $|i, j|$ is:

$$|i, j| = \sqrt{|x_i - x|^2 + |y_i - y|^2} \tag{2}$$

Through formula (1) and formula (2), it is possible to judge whether any node i can be successfully automatically identified when abnormal. If the automatic identification probability is 1 for all nodes in the wireless sensor network after the operation of formula (1), it means that the entire wireless sensor network is perceivable, and the communication data of each node in the wireless sensor network can be obtained. Otherwise, it indicates that there are coverage blind spots in the network, and it is necessary to partition the network and arrange automatic identification nodes in each partition to ensure that the entire wireless sensor network maintains a perceivable state and realize the acquisition of wireless sensor network communication data under the damaged state.

The above node identifiability model provides theoretical criteria for evaluating the global perceptibility of the network. To apply it to actual data collection, this study instantiated the model as follows: the physical coordinates of nodes and central nodes were derived from pre deployment network topology mapping; The optimal perception radius is determined through field link measurement, which tests the packet reception rate of the central node and surrounding nodes at different distances in the deployment environment. The farthest reliable communication distance that meets the minimum communication quality requirements is defined as this radius. In this experimental environment, the value of each cluster head node is distributed between 80 meters and 120 meters. The recognition probability is integrated into network management software as a binary decision function in practice, which determines whether a node can be monitored by calculating the distance between nodes in real-time and comparing it with the corresponding optimal perception radius. In this experimental deployment, by optimizing the node layout, the recognition probability of all 150 nodes was ensured to be 1, thus meeting the prerequisite of "global perceptibility" and providing a complete data foundation for subsequent analysis. Based on this network, all communication data is collected through the actual protocol stack to form the initial high-dimensional dataset; The simulation of network damage state is achieved through software injection of damage on this normal data stream, in order to obtain a controllable dataset of damage state.

In the actual data collection process, this study is based on the above model to guide network deployment and optimization, ensuring that the global perceptible conditions are met as much as possible in the actual physical topology. All communication data (including normal and abnormal traffic) are collected through this actual network, and the data collection process follows standard communication protocols, recording multi-dimensional information such as communication

timing, packet content, and signal strength of each node, forming the original high-dimensional dataset for subsequent dimensionality reduction and feature extraction.

2.2 Data dimensionality reduction of wireless sensor network communication network under damaged state based on kernel principal component analysis

The wireless sensor network networking model can achieve global state awareness of wireless sensing. Obtaining the communication data of each node in the wireless sensor network not only involves a huge amount of data, but also has a very high data dimension. In this state, data transmission will consume a large amount of communication resources. However, due to factors such as packet loss, channel congestion, and environmental interference in the network damaged state, it will cause great difficulties in the communication of the wireless sensor network, making it difficult to transmit data. Therefore, it is necessary to reduce the dimension of the communication data of the wireless sensor network in the network damaged state. By reducing the data dimension of the wireless sensor network in the damaged state, reducing data complexity, improving data calculation speed, occupying less communication resources in the damaged state, and improving the communication anomaly intrusion detection speed of the wireless sensor network in the network damaged state [20]. In order to quickly and accurately reduce the dimension, the kernel method is introduced based on the principal component analysis method to form the kernel principal component analysis method for data reduction [21].

The basic principle of the kernel method is as follows: The communication data of the wireless sensor network in the damaged state in the input data space is mapped to a high-dimensional feature space through a non-linear function, and data processing is carried out in the feature space. In this process, a kernel function is introduced to convert the inner product operation in the feature space after non-linear transformation into the calculation of the kernel function in the original space, thereby reducing the computational amount. The process of using kernel principal component analysis for dimensionality reduction of wireless sensor network communication data is shown in Figure 1.

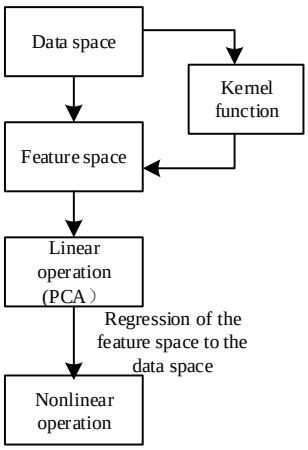


Figure 1: Kernel method framework

Based on the kernel method framework shown in Figure 1, the KPCA dimensionality reduction process achieves nonlinear feature extraction of high-dimensional data through four key stages: in the kernel function mapping stage, the original high-dimensional data is projected into a high-dimensional feature space through a nonlinear mapping function. This process uses kernel functions to implicitly calculate the inner product of samples in the feature space, avoiding complex high-dimensional explicit calculations. In the linear operation stage, data in the feature space is processed through standard principal component analysis. By calculating the eigenvalues and eigenvectors of the covariance matrix, determine the direction of maximum data variance and establish a low dimensional orthogonal coordinate system. In the feature space data reconstruction stage, the projection coordinates of the original data on the principal components of the feature space are obtained through mathematical transformations. These projection coordinates form the reduced dimensional dataset, preserving the key nonlinear features of the original data. Finally, through the organic combination of the above stages, nonlinear dimension reduction was achieved, resulting in a low dimensional data representation with discriminative power, providing an effective data foundation for subsequent anomaly intrusion detection. The specific process is as follows:

The nuclear method is combined with the principal component analysis to form the kernel principal component analysis method. The wireless sensor network communication obtained in Section 2.1 above. The trusted data is set as Π wireless sensor network communication data samples $\pi_k \left(k=1,2,3,\cdots,\Pi\right)$ in the input data space, $\pi_k \in R^N$, such that $\sum_{k=1}^{\Pi} \pi_k = 0$, and its covariance matrix $\boldsymbol{C}$ is:

$$\boldsymbol{C} = \frac{\sum_{k=1}^{\Pi} \pi_k \pi_k^T}{\Pi} \quad (3)$$

In the formula, T is the transpose.

In the PCA algorithm, a non - linear mapping function ϕ is introduced to transform the communication data sample points $\pi_1,\pi_2,\pi_3,\cdots,\pi_\Pi$ of the wireless sensor network in the damaged state in the input data space into the communication data sample points $\phi(\pi_1),\phi(\pi_2),\phi(\pi_3),\cdots,\phi(\pi_\Pi)$ of the wireless sensor network in the damaged state in the feature space. Assume:

$$\sum_{k=1}^{\Pi}\phi(\pi_k)=0 \qquad (4)$$

In the feature space G, the covariance matrix is:

$$\bar{\boldsymbol{C}}=\frac{\sum_{k=1}^{\Pi}\phi(\pi_k)\phi(\pi_k)^T}{\Pi} \qquad (5)$$

Therefore, the formula for solving the communication data value and vector of the wireless sensor network in the feature space G is:

$$\gamma\boldsymbol{\nu}=\bar{\boldsymbol{C}}\boldsymbol{\nu},\boldsymbol{\nu}\in G\neq\{0\} \qquad (6)$$

Where, γ and $\boldsymbol{\nu}$ are the wireless sensor network communication data values and vectors in the feature space G, respectively.

Thus, it can be obtained that:

$$\gamma\left(\phi(\pi_k)\boldsymbol{\nu}\right)=\phi(\pi_k)\bar{\boldsymbol{C}}\boldsymbol{\nu} \qquad (7)$$

The linear representation formula of the communication data vector $\boldsymbol{\nu}$ of the wireless sensor network in the feature space G is:

$$\boldsymbol{\nu}=\sum_{i=1}^{\Pi}\boldsymbol{\delta}_i\phi(\pi_i) \qquad (8)$$

In the equation, $\boldsymbol{\delta}_i$ is the eigenvector.

According to formulas (5), (7), and (8), we can obtain:

$$\gamma\sum_{\substack{i=1\\k=(1,2,3,\cdots,\Pi)}}^{\Pi}\boldsymbol{\delta}_i\left(\phi(\pi_k)\cdot\phi(\pi_i)\right)=\frac{\sum_{i=1}^{\Pi}\boldsymbol{\delta}_i\left(\phi(\pi_k)\cdot\sum_{i=1}^{\Pi}\phi(\pi_k)\right)\left(\phi(\pi_j)\cdot\phi(\pi_i)\right)}{\Pi} \qquad (9)$$

Define the matrix $\boldsymbol{K}$ of $\Pi\times\Pi$:

$$\boldsymbol{K}_{i,j}=\phi(\pi_i)\cdot\phi(\pi_j) \qquad (10)$$

Then formula (9) can be transformed into:

$$\Pi\gamma\boldsymbol{\delta}=\boldsymbol{K}\boldsymbol{\delta} \qquad (11)$$

By solving formula (11), the communication data value γ and vector $\boldsymbol{\nu}$ of the wireless sensor network in the mapping space can be obtained. Arrange $\gamma_1,\gamma_2,\gamma_3,\cdots,\gamma_n$ in descending order and adjust $\boldsymbol{\nu}_1,\boldsymbol{\nu}_2,\boldsymbol{\nu}_3,\cdots,\boldsymbol{\nu}_n$ to correspond to the sorted $\gamma_1,\gamma_2,\gamma_3,\cdots,\gamma_n$. Using the Gram Schmidt orthogonal method to normalize vectors, ensuring that the extracted principal components are orthogonal to each other, thus constructing a non redundant and discriminative feature representation in a low dimensional space, and obtaining $\boldsymbol{\delta}_1,\boldsymbol{\delta}_2,\boldsymbol{\delta}_3,\cdots,\boldsymbol{\delta}_n$. Then the projection of the communication data test sample of the wireless sensor network on the vector $\boldsymbol{\nu}^k$ in the feature space G is:

$$y=\sum_{i=1}^{n}\boldsymbol{\delta}_i\,\boldsymbol{K}\left(\pi_i,\pi_j\right) \qquad (12)$$

If the communication data of the wireless sensor network in the feature space does not meet the centering condition, the matrix needs to be corrected. Replace $\boldsymbol{K}$ in formula (12) with $\bar{\boldsymbol{K}}$, and the formula is:

$$\bar{\boldsymbol{K}}_{i,j}=\boldsymbol{K}_{i,j}-\frac{\sum_{m=1}^{\Pi}o_{i\Pi}\boldsymbol{K}_{\Pi j}}{\Pi}-\frac{\sum_{n=1}^{\Pi}o_{in}\boldsymbol{K}_{nj}}{\Pi}+\frac{\sum_{\Pi,n=1}^{\Pi}o_{im}\boldsymbol{K}_{mn}o_{nj}}{\Pi^2} \qquad (13)$$

In formula (13), i , $j=1$, and $o_{i,j}$ are correction coefficients.

The specific process of dimensionality reduction of the communication data of the wireless sensor network by the kernel principal component analysis method according to the above content is as follows:

(1) Suppose Π data records are obtained from the communication data of the wireless sensor network sensed by the wireless sensor network networking model (each record has n attribute components), and represent it as a $\Pi\times n$ -dimensional matrix:

$$\boldsymbol{\pi}_k = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1n} \\ \vdots & \ddots & \vdots \\ \pi_{\Pi 1} & \cdots & \pi_{\Pi n} \end{bmatrix} \quad (14)$$

(2) Select an appropriate kernel function and calculate the kernel matrix $\boldsymbol{K}$. Since the Gaussian radial basis function (RBF) kernel function can implicitly map data to a high-dimensional space to achieve nonlinear modeling, with few parameters, strong adaptability, and relatively efficient calculation, the Gaussian radial basis function is selected as the kernel function of kernel principal component analysis. The formula of the Gaussian radial basis function is:

$$\boldsymbol{K}\left(\pi_i, \pi_j\right) = \exp\left(-\frac{\left\|\pi_i - \pi_j\right\|^2}{\sigma^2}\right) \quad (15)$$

In the formula, $\left\|\pi_i - \pi_j\right\|^2$ is the square of the Euclidean distance between two communication data of the wireless sensor network in the feature space; σ is the kernel function width.

(3) Correct the kernel matrix $\boldsymbol{K}$ to obtain $\bar{\boldsymbol{K}}$, as shown in formula (13).

(4) Calculate the value $\gamma_1, \gamma_2, \gamma_3, \cdots, \gamma_n$ of $\bar{\boldsymbol{K}}$ and the vector $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3, \cdots, \boldsymbol{\nu}_n$.

(5) Arrange $\gamma_1, \gamma_2, \gamma_3, \cdots, \gamma_n$ in descending order and adjust $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3, \cdots, \boldsymbol{\nu}_n$ to correspond to the sorted $\gamma_1, \gamma_2, \gamma_3, \cdots, \gamma_n$.

(6) Use the Gram-Schmidt orthogonal method to normalize the vector and obtain $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3, \cdots, \boldsymbol{\delta}_n$.

(7) Calculate the cumulative contribution rate $B_1, B_2, B_3, \cdots, B_n$ of the sorted $\gamma_1, \gamma_2, \gamma_3, \cdots, \gamma_n$ values. According to the given extraction efficiency p , if $B_i \geq \rho$, then extract i principal components $\delta_1, \delta_2, \delta_3, \cdots, \delta_i$.

(8) Calculate the projection $y = \bar{\boldsymbol{K}}\boldsymbol{\delta}$ of the corrected kernel matrix $\bar{\boldsymbol{K}}$ on the extracted corresponding vectors, where $\boldsymbol{\delta} = \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3, \cdots, \boldsymbol{\delta}_i$; the obtained projection y is the data obtained after the data is reduced in dimension by KPCA.

After reducing the dimension of the wireless sensor network communication data in the network damaged state by the KPCA method, a certain degree of compression is performed, redundant information is removed, and the complexity of the data is reduced, so that the data can be transmitted using fewer communication resources. Occupying fewer communication resources in the network damaged state can transmit data faster, so the wireless sensor network communication anomaly intrusion detection can be completed faster.

2.3 Feature extraction of wireless sensor network communication data in the damaged state based on the restricted boltzmann machine

After completing the dimension reduction of the wireless sensor network communication data in the network damaged state by the KPCA method, the Restricted Boltzmann Machine (RBM) is used to extract the data features of the data after dimension reduction in the network damaged state. RBM is a probability distribution function based on energy [22], and its energy function $E(\mathrm{y}, h)$ formula is:

$$E(\mathrm{y}, h) = -\sum_{i=1}^{n_y} \alpha_i y_i - \sum_{j=1}^{n_h} \beta_j y_j - \sum_{i=1}^{n_y}\sum_{j=1}^{n_h} h_j \boldsymbol{\omega}_{j,i} y_i \quad (16)$$

Where: $h = \left(h_1, h_2, h_3, \cdots, h_m\right)^T$ represents the values of the neurons in the RBM hidden layer, $y = \left(y_1, y_2, y_3, \cdots, y_n\right)^T$ represents the values of the neurons in the visible layer, that is, the wireless sensor network communication data after dimension reduction, α_i is the bias value of the visible layer neuron i , β_j is the bias value of the hidden layer neuron j , and $\boldsymbol{W} = \left(\boldsymbol{\omega}_{ij}\right)_{m \times n}$ is the connection weight matrix from the visible layer to the hidden layer.

According to the energy function, the joint probability distribution $P(y, h)$ of y and h can be obtained as:

$$P(y, h) = \frac{1}{Z} e^{-E(y,h)} \quad (17)$$

Where, Z is the partition function (normalization factor):

$$Z = \sum_{y,h} e^{-E(y,h)} \qquad (18)$$

Formula (17) gives the "energy - probability" relationship, from which the feature μ of the wireless sensor network communication data after dimension reduction extracted by the hidden layer can be obtained, and the formula is:

$$\mu = P(y_i = 1 \mid h) = \sigma\left(\alpha_j + \sum_i h_j \omega_{ij}\right) \qquad (19)$$

In the equation, σ is the Sigmoid function.

According to formula (20), the dimensionality-reduced wireless sensor network communication data features μ extracted by the hidden layer of the RBM [23], that is, the communication features of the wireless sensor network in the network damaged state.

When the RBM algorithm is applied to the process of extracting wireless sensor network communication data features, an unsupervised learning method is adopted to train the RBM feature extraction model. The purpose of training the RBM is to make the distribution $p(y)$ of the visible layer nodes y best fit the distribution $q(y)$ of the sample space where the input dimensionality-reduced wireless sensor network data samples are located. From the perspective of information entropy, it is to make the *KL* (Kullback-Leibler, relative entropy) distance between p and q the smallest, so that the two distributions are closer. The formula is:

$$\begin{aligned} KL(q \| p) &= \sum_{y \in Y} \ln \frac{q(y)}{p(y)} \\ &= \sum_{y \in Y} q(y) \ln q(y) - \sum_{y \in Y} q(y) \ln p(y) \end{aligned} \qquad (20)$$

Since the input of the visible layer is the communication data information in the wireless sensor network, $q(y)$ is a determined item. To ensure the smallest *KL* distance, $\sum_{y \in Y} q(y) \ln p(y)$ needs to be maximized. Since the sample space Y of the wireless sensor network communication data is unknown, the Monte Carlo method is used to find the approximate value of $\sum_{y \in Y} q(y) \ln p(y)$. The formula is:

$$\sum_{y \in Y} q(y) \ln p(y) \approx \frac{\sum_{\psi=1}^{\Psi} \ln p\left(y^{\psi}\right)}{\Psi} \qquad (21)$$

In the formula: Ψ is the assumed number of training samples; y^{ψ} represents the ψ th training sample.

To find the optimal RBM parameter $\theta = \{W, \alpha, \beta\}$, the logarithmic loss function is adopted. The formula is:

$$Loss = -\sum_{i=1}^{\psi} \ln p\left(y^{(i)}\right) \qquad (22)$$

Thus, the training of the RBM feature extraction model is completed, and the feature extraction of the dimensionality-reduced wireless sensor network communication data is realized based on the trained RBM.

To ensure the effectiveness and reproducibility of the RBM feature extraction model, this study specifically designed its network structure and training process. In terms of network structure, the number of visible layer neurons remains consistent with the dimensionality of KPCA reduced data; The number of hidden layer neurons has been experimentally verified to be 128, in order to achieve a balance between feature compression and information preservation. The model training adopts the Contrastive Divergence (CD-1) algorithm, which performs one-step Monte Carlo simulation through Gibbs sampling to efficiently approximate the negative phase gradient of the data. Specifically, during the training process, Gibbs sampling is used to quickly generate visible and hidden layer samples from the current distribution of RBM in order to approximate the expected term of the model in formula (21). This Monte Carlo sampling method avoids the huge computational overhead of directly calculating the partition function, and approximates the true gradient direction through finite step sampling, making it possible to maximize the logarithmic likelihood function through the random gradient ascent method. The key hyperparameters are set as follows: learning rate of 0.01, training epochs of 100, and batch size of 64. This set of parameter combinations has been experimentally verified to effectively ensure the stable convergence of the model and extract deep features with high discriminative power.

3 Abnormal intrusion detection of wireless sensor network communication in the damaged state based on Naive Bayes Classifier

The naive Bayes classifier ensures that the posterior probability of a class is higher than that of other classes according to the maximum posterior probability rule, so as to achieve accurate classification [24]. The extracted wireless sensor network communication data feature vector $\boldsymbol{\mu}$ is used as the input of the naive Bayes network for abnormal intrusion detection of wireless sensor network communication [25]. Since the normal wireless sensor network in the network damaged state will also be affected by different factors, resulting in certain communication anomalies, but there are differences in the performance of the abnormal feature vector $\boldsymbol{\mu}$ between the communication anomalies caused by the network damaged state and those caused by intrusions. Therefore, the abnormal intrusion detection of communication in the network damaged state can be realized through the abnormal feature vector $\boldsymbol{\mu}$.

First, set or adjust the weight vector $\boldsymbol{\varepsilon}$ of the prior probability. The formula is:

$$p\left(\boldsymbol{\varepsilon}\mid\kappa\right)=\prod_{1}^{\vartheta}N\left(0,\partial_{i}^{-1}\right) \qquad (23)$$

In the formula: ϑ is the hyperparameter vector described by ∂.

Use $S=\left[s_1,s_2,s_3,\cdots,s_n\right]$ to describe the set of wireless sensor network communication data. Use $\boldsymbol{U}=\left[\boldsymbol{\mu}_2,\boldsymbol{\mu}_2,\boldsymbol{\mu}_3,\cdots,\boldsymbol{\mu}_n\right]$ to describe the set of wireless sensor network communication data feature vectors. Adopt the $s=f\left(\boldsymbol{\mu}_i\right)$ mapping criterion for description, where f is a classifier, which can make a random $s_i\in S$ uniquely correspond to a $\boldsymbol{\mu}_i\in\boldsymbol{U}$, satisfying $s=f\left(\boldsymbol{\mu}_i\right)$. The naive Bayes classification method is to establish a naive Bayes classifier, and the probability $p\left(s_e\mid\boldsymbol{\mu}\right)$ that the set of homogeneous data feature vectors $\boldsymbol{\mu}$ of wireless sensor network nodes that need to be classified as normal or abnormal belongs to the s_e class. The formula is:

$$p\left(s_e\mid\boldsymbol{\mu}\right)=\frac{p\left(\boldsymbol{\mu}\mid s_e\right)p\left(s_e\right)}{p\left(\boldsymbol{\mu}\right)} \qquad (24)$$

In the formula: $p\left(\boldsymbol{\mu}\right)>0$; $p\left(s_e\right)>0$, $e=1,2,3,\cdots,m$.

In formula (24), for different wireless sensor network states, the denominator $p\left(\boldsymbol{\mu}\right)$ is fixed, so maximizing the numerator is sufficient. If the wireless sensor network communication data in the network damaged state in the training set follows a specific probability distribution or is parameter-free, when dealing with continuous data, it is assumed that this data follows a normal distribution. For a certain continuous attribute in the training set, represented by b , first classify it, and then calculate the mean and variance of each class. Use η_{s_1} to represent the variance of b in the s_1 class. The communication abnormal intrusion probability $p\left(\boldsymbol{\mu}=z\mid s_1\right)$ for any wireless sensor network node is:

$$p\left(\boldsymbol{\mu}=z\mid s_1\right)=\frac{p\left(s_e\boldsymbol{\mu}\right)e}{p\left(\boldsymbol{\varepsilon}\middle|k\right)\sqrt{2\pi\eta_{s_1}^2}} \qquad (25)$$

In the formula, η_{s_1}、z are respectively used to describe the normal distribution and the variance with the mean at s_1, $p\left(\boldsymbol{\varepsilon}\mid\kappa\right)$ is a naive Bayes classifier.

Since normal network communication in the network damaged state will also be interfered and generate certain anomalies, a threshold is set based on formula (25), adjusted according to the actual situation based on a 60% probability. When the abnormal probability exceeds this threshold, it is determined that there is a communication abnormal intrusion in this wireless sensor network, and there is a node being invaded in the wireless sensor network.

Since the network damaged state will bring great interference to the intrusion detection of the naive Bayes classifier, the features of the naive Bayes classifier are weighted and the micro conditional probability is optimized.

(1) Feature Weighting Optimization

The formula of the naive Bayes classifier based on feature weighting optimization is:

$$p\left(\boldsymbol{\varepsilon}\mid\kappa\right)=\arg\max p\left(\boldsymbol{\varepsilon}\right)\prod_{1}^{\vartheta}N\left(\varsigma_i\mid\boldsymbol{\varepsilon}\right)^{\omega_\tau}\prod_{1}^{\vartheta}N\left(\varsigma_i\mid\boldsymbol{\varepsilon}\right)^{\omega_t} \qquad (26)$$

In the formula, the weight vectors $\boldsymbol{\varepsilon}=\left\{\omega_\tau,\omega_t\right\}$, ω_τ is the qualitative weight calculated based on the mutual information between features and categories, reflecting the correlation between features and abnormal intrusion. ω_t is the quantitative weight obtained based on the

inverse normalization of feature variance, used to balance the impact of differences in different feature value ranges. The specific calculation of weights is determined on the training set through cross validation.

(2) Micro Conditional Probability Optimization

In the training phase, first learn the estimated information and the Naive Bayes classifier, and then predict the training set to obtain the misclassified training samples. In the fine-tuning phase, modify the conditional probabilities corresponding to all misclassified samples in each round. Increase the conditional probability of each feature under the true class by a certain step size and decrease the conditional probability of each feature under the predicted class by a certain step size, so as to improve the accuracy of classifying the training set in the next round. Otherwise, stop the iteration. The optimal conditional probability is the conditional probability modified in the previous round. The formula is:

$$p(s_e \mid \boldsymbol{\mu}) = \frac{p(\boldsymbol{\mu} \mid s_e)\, p(s_e)}{p(\boldsymbol{\mu})} + \varpi_t(TC)$$
$$p(s_e \mid \boldsymbol{\mu}) = \frac{p(\boldsymbol{\mu} \mid s_e)\, p(s_e)}{p(\boldsymbol{\mu})} + \varpi_t(PC) \quad (27)$$

Among them, TC and PC are the true class and the predicted class of the training samples of the communication data of the wireless sensor network respectively. ϖ_t is the conditional probability of the t -th round of iteration, and t is the number of iterations.

In the fine-tuning phase, an iterative optimization strategy is adopted to adjust the conditional probability estimates, specifically implemented through the following algorithm:

Input: Training dataset, initial Naive Bayes classifier, maximum iteration count *T*, convergence threshold ε

Output:Optimized conditional probability table

1)Initialization:

Set the current iteration round t=0, train using the training dataset to obtain the initial conditional probability table.

2)Iterative Optimization:

for t=1 to T do

Step 1: Use the current conditional probability table to predict the training set and collect the set of misclassified samples.

Step 2: Check convergence criteria:

if the number of misclassified samples stops decreasing and the decrease magnitude is less than threshold ε then

Terminate the iteration, jump to Step 5.

end if

Step 3: For each misclassified sample:

Obtain the sample's true class and predicted class.

for each feature do

Increase the conditional probability of this feature under the true class (increase by step size δ).

Decrease the conditional probability of this feature under the predicted class (decrease by step size δ).

end for

Step 4: Perform normalization to ensure the sum of all conditional probabilities is 1.

end for

3)Output Result:

Step 5: Return the final optimized conditional probability table.

The core idea of this optimization process is: for each misclassified sample, increase the conditional probabilities of its features under the true class, while simultaneously decreasing them under the incorrectly predicted class. Through this bidirectional adjustment, the model can gradually correct initial estimation biases and better adapt to the true data distribution.

During optimization, the adjustment step size δ for conditional probabilities is typically set to a small positive number, such as 0.01 or 0.05, to ensure a smooth and convergent optimization process. Additionally, the algorithm incorporates dual convergence criteria: firstly, the number of misclassified samples ceases to decrease, and secondly, the decrease magnitude falls below a preset threshold. These two conditions together ensure the optimization process terminates at an appropriate point, preventing overfitting.

By performing feature weighting on the Naive Bayes classifier and optimizing the conditional probabilities through fine-tuning, the influence of key features on classification can be highlighted, the weights of features with a high degree of association with the target category can be enhanced, the interference of irrelevant features can be suppressed, and at the same time, the probability estimation deviation caused by assuming feature independence, data sparsity, etc. can be corrected, making the model more adaptable to the actual distribution of the data, and increasing the accuracy of detecting abnormal intrusions in the wireless sensor network communication under the network damaged state.

4 Experimental analysis

4.1 Experimental object

The experiment selected a wireless sensor network used in a company, which covers over 1000 square meters of production, warehousing, and transportation areas, with a total of 150 sensor nodes covering various types such as vibration, temperature and humidity, pressure, etc. It is used for real-time monitoring of equipment production and environmental parameters for intelligent control. The topology structure of the wireless sensor network is shown in Figure 2.

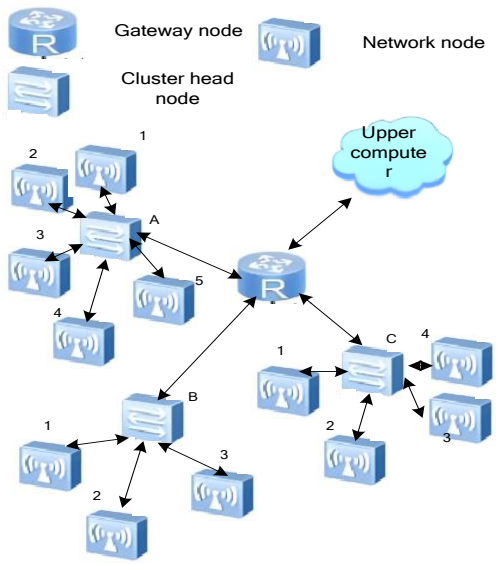


Figure 2: Topology structure of wireless sensor network

Figure 2 illustrates the topology structure of the Wireless Sensor Network (WSN) employed in the experiments. This network is deployed in an industrial environment, covering production, storage, and transportation areas for real-time monitoring of equipment operating status and environmental parameters. A hybrid topological structure is adopted, combining the advantages of star and mesh topologies to enhance communication reliability and coverage. The network comprises multiple cluster head nodes (CHs) (e.g., Cluster Head A, B, C), which are responsible for coordinating data aggregation and transmission from ordinary sensor nodes within their respective regions. Utilizing a multi-hop communication mechanism, data is relayed through the cluster head nodes to a gateway node, and ultimately uploaded to the cloud or a control center. To verify the anomaly intrusion detection capability of the proposed method under network damage conditions, communication damage was artificially introduced in the experiment: a random packet loss rate of 30% was set between cluster head C and the gateway to simulate channel quality deterioration; Limit the communication bandwidth of cluster heads A and B to 40% of their original value, simulating a resource constrained scenario. The dataset used for training and testing was actually collected through the network, containing a total of 50000 communication data records, including 35000 normal communication data and 15000 abnormal data. Abnormal data is generated by simulating ten typical attack behaviors, including Sinkhole attack, wormhole attack, denial of service attack, replay attack, disguised node attack, selective forwarding attack, flooding attack, data tampering attack, key leakage attack, and topology destruction attack. Each type of attack generates approximately 1500 instances to ensure class balance in the dataset, and each type of attack is injected into specific nodes through scripts, with corresponding timestamps and traffic characteristics recorded. The ratio of normal and abnormal samples in the dataset is 7:3, ensuring that the model still has good generalization ability even in cases of class imbalance. All data are anonymized during the collection process and feature sequences are constructed using time window slicing to support subsequent dimensionality reduction and feature extraction operations.

The parameters of the wireless sensor network used are shown in Table 2.

Table 2: Parameters of wireless sensor network

Attribute	Parameters
Topological structure	Hybrid topological structure
Communication standard	IEEE 802.15.4g
Working frequency band	2.4GHz
Network capacity	Support concurrent access of over 2000 nodes
Sensor type	Vibration sensor, temperature and humidity sensor, RFID reading and writing module, pressure sensor
Battery life	Lithium thionyl chloride battery, with a 5-year battery life (in low-power mode)
Protection grade	IP67
Computing power	8-bit microcontroller, supporting edge lightweight computing
Transmission distance	Single jump 150 meters
Data rate	250kbps
Encryption protocol	AES - 128
Deployment method	Guide rail installation + ceiling mounting + pole clamping installation

Gateway interface	Ethernet (1Gbps), 4G/5G modules
Working temperature	-20℃ ~ 70℃

The selection of kernel functions and their parameters has a decisive impact on the dimensionality reduction performance of KPCA. This study uses Gaussian radial basis function (RBF) kernel function, whose parameter kernel width σ controls the flexibility and locality of the model. To evaluate the robustness of parameter selection and determine the optimal value, a sensitivity analysis experiment was designed: on the same dataset, all other conditions were fixed, and the value of σ was systematically changed to observe its impact on the performance of reduced dimensional data in subsequent anomaly intrusion detection tasks (with F1 score as the core indicator).

Table 3: Sensitivity analysis results of KPCA parameter σ

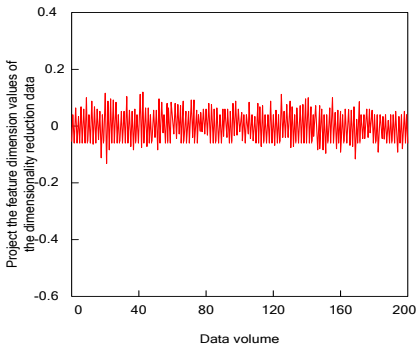
Kernel width	Data dimension after dimensionality reduction	Abnormal intrusion detection F1 score
0.01	28	0.71
0.10	19	0.84
0.50	15	0.92
1.00	13	0.87
5.00	10	0.78

According to the results in Table 3, the performance of KPCA is more sensitive to changes in the value of σ. When σ =0.50, the reduced dimensional data achieved the highest F1 score (0.92) in subsequent classification tasks, indicating that under this parameter, KPCA can most effectively extract the nonlinear features that are most beneficial for distinguishing normal and abnormal communication modes. Meanwhile, when σ is within the range of [0.10, 1.00], the F1 scores remain above 0.84, indicating that the method has good robustness within this parameter range. All subsequent experiments in this study were conducted using the optimal parameter determined by this experiment, σ =0.50.

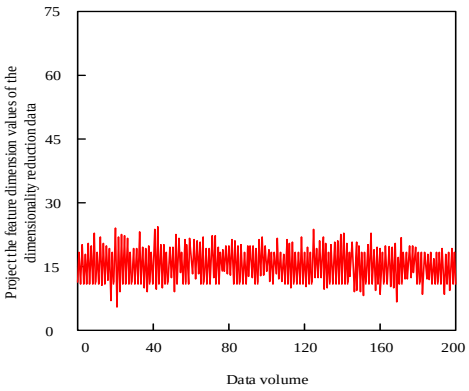
To ensure the reproducibility and scientificity of the proposed method, this study specifically set and explained the hyperparameters of the model. Firstly, in the design of a Restricted Boltzmann Machine (RBM), the number of hidden layer neurons is set to 128 to achieve a balance between feature compression and information preservation; The model training uses the Contrastive Divergence (CD-1) algorithm, with key hyperparameters including a learning rate of 0.01, 150 training epochs, and a batch size of 64. Secondly, in the process of optimizing the Naive Bayes classifier, the qualitative weights (based on mutual information) and quantitative weights (based on reciprocal variance) used for feature weighting are fused with coefficients of 0.7 and 0.3, respectively, and determined through cross validation; The adjustment step size for micro conditional probability optimization is set to 0.01, the maximum number of iterations is 50, and the convergence criterion is to stabilize the number of misclassified samples (with a change amplitude less than 0.001). In the process of determining key parameters of the model, such as the kernel width σ of KPCA, the number of hidden units in RBM, and the weight coefficients of Naive Bayes, 5-fold cross validation was used to optimize on the training set. Specifically, divide the training set into 5 equal parts, take turns using 4 parts as the training subset and the remaining 1 part as the validation subset, and loop 5 times. The final parameter selection is based on the group with the best average performance among 5 verifications. All performance metrics reported, such as F1 score and false positive rate, are the final results of the model on an independent test set.

4.2 Analysis of the effectiveness of the proposed method

To verify the effect of the KPCA method used in this article on dimensionality reduction of original wireless sensor network communication data, wireless sensor network communication data was randomly collected and dimensionality reduction was performed using the KPCA method proposed in this article. The original wireless sensor network communication dataset contains 90 features (attributes), and the reduced wireless sensor network communication data is shown in Figure 3.



(a)The distribution of original wireless sensor network communication data



(b)Data distribution of wireless sensor network communication after KPCA dimension reduction

Figure 3: Data dimensionality reduction effect of the KPCA method

From Figure 3, it can be seen that the KPCA method proposed in this paper can achieve efficient dimensionality reduction of wireless sensor network communication data. In the original state, the feature dimensions of wireless sensor network communication data without dimensionality reduction are mainly concentrated in 60-80 features, but there are also a large number of fluctuations between 10-90 features, resulting in extremely high data dimensions. In the state of network damage, the computing power of the network decreases, and it is difficult to ensure the accuracy of anomaly intrusion detection by processing high-dimensional data. After using the KPCA method in this article for data dimensionality reduction (as shown in Figure 3 (b)), the projected values of the data in the KPCA principal component space were standardized to around 15 features, which cumulatively explained about 94.7% of the original variance, indicating that the reduced features had been effectively compressed and concentrated, with slight fluctuations but almost negligible. The data fluctuation is reduced and the dimensionality is significantly reduced after dimensionality reduction using the method described in this article. In situations where the network is damaged and communication resources are tight, very few resources can be used for anomaly intrusion detection to ensure real-time detection. Moreover, the reduced dimensional data reduces interference, improves the accuracy of anomaly intrusion detection, and ensures that the proposed method can quickly and accurately detect communication anomalies in wireless sensor networks, thereby ensuring the security of wireless sensor networks.

The data fluctuates less after dimensionality reduction by this method, and the dimension is significantly reduced. In the case of network damage and tight communication resources, extremely few resources can be used for anomaly intrusion detection, ensuring the real - time performance of detection. Moreover, the dimensionality - reduced data reduces interference and improves the accuracy of anomaly intrusion detection, ensuring that the method in this paper can quickly and accurately detect the abnormal intrusion of wireless sensor network communication, thus ensuring the security of the wireless sensor network.

To quantitatively evaluate the data dimensionality reduction performance of the kernel principal component analysis (KPCA) method used in this paper, we compared it with two widely used nonlinear dimensionality reduction methods - t-distributed random neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP). The evaluation metric is the cumulative variance contribution rate, which reflects the ability of the reduced data to retain the original data information and is a key quantitative standard for measuring the effectiveness of dimensionality reduction. The cumulative variance contribution rates of the three methods in extracting different numbers of principal components are shown in Table 4.

Table 4: Cumulative variance contribution rate of different dimensionality reduction methods (%)

Number of Principal Components	KPCA	t-SNE	UMAP
5	68.5	45.2	52.7
10	86.3	61.8	70.4
15	94.7	73.1	81.9
20	98.2	80.5	88.3
25	99.5	85.0	92.1

The following conclusion can be drawn from Table 3: The KPCA method used in this article is significantly better than the comparative method in terms of cumulative variance contribution rate. Specifically, with only 15 principal components, KPCA can retain 94.7% of the variance information in the original data; However, t-SNE and UMAP can only retain 73.1% and 81.9% of the information under the same principal component score. To achieve information retention levels similar to KPCA (approximately 94%), t-SNE and UMAP require more dimensions, which undoubtedly increases the burden of subsequent computation and communication. This result demonstrates that KPCA can capture key nonlinear structures in wireless sensor network communication data with fewer dimensions and higher

efficiency, and its dimensionality reduction performance is superior to t-SNE and UMAP. This advantage stems from the ability of kernel methods to implicitly handle inner product operations in high-dimensional feature spaces, avoiding the computational complexity of direct high-dimensional mapping while preserving key structural information of the data. In the event of network damage, KPCA not only reduces the dimensionality of the data, but more importantly, extracts more discriminative features through nonlinear mapping, providing higher quality and less noisy input data for subsequent anomaly intrusion detection, thereby improving the overall robustness and detection efficiency of the system.

4.3 Experimental comparison results

To verify the effectiveness of the RBM method used in this paper for extracting reduced dimensional wireless sensor network features, we compared our method with three feature extraction methods: CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory Neural Network), and LLE (Local Linear Embedding). The features extracted by the above methods were input into a Bayesian classifier for communication anomaly intrusion detection. To ensure fairness in comparison, CNN adopts a typical structure consisting of two convolutional layers (kernel sizes of 3 and 5 respectively) and one fully connected layer; LSTM uses a single-layer network with 128 hidden units to capture temporal dependencies. All models use the Adam optimizer and learn until convergence on the same training set. All the features extracted by the methods were input into the same Naive Bayes classifier for communication anomaly intrusion detection, and the F1 score was used as the evaluation metric. The results are shown in Figure 4.

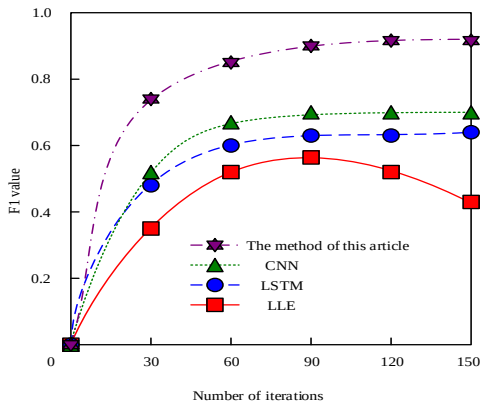


Figure 4: F1 values of different feature extraction methods under 75% network damage rate

From Figure 4, it can be seen that among the different feature extraction methods mentioned above, the RBM method proposed in this paper combines Naive Bayes classifier for anomaly intrusion detection after feature extraction, with an F1 value of 0.9, significantly higher than other methods. This indicates that the RBM method can more effectively extract key features from wireless sensor network communication data, thereby improving the accuracy of classification. This is mainly due to the fact that RBM, as an energy based unsupervised deep learning model, can effectively capture nonlinear feature distributions in high-dimensional data, especially suitable for complex and nonlinear abnormal patterns caused by network damage in wireless sensor networks. WSN communication data usually has the characteristics of high dimensionality, strong noise, and pattern mutation. The probability generation model of RBM can learn its underlying distribution well and is insensitive to noise, thus robustly extracting key features. In contrast, although CNN performs well in image and sequence data processing, its deep structure is prone to overfitting in the small sample, high noise wireless sensor network communication data used in this experiment, and the assumption of local correlation in the data may not always hold true in this task, resulting in limited feature extraction ability, with an F1 value of only about 0.6. LSTM is also difficult to learn effective long-term time dependencies due to limited data volume and noise interference, and its F1 value is similar to CNN. The LLE method has the lowest F1 value after feature extraction, and there is a decrease in F1 value during the iteration process. This may be due to the poor performance of LLE methods in processing high-dimensional and nonlinear data, resulting in extracted features that cannot effectively distinguish between normal and abnormal communication data. The above results indicate that the proposed RBM feature extraction method has significant advantages in wireless sensor network communication anomaly intrusion detection, which can more effectively extract key features and improve classification accuracy.

In order to verify the effectiveness of intrusion detection for abnormal communication in wireless sensor networks under network damage conditions, the company's wireless sensor network topology caused a certain amount of packet loss in the communication between cluster head C and gateway nodes, and reduced the bandwidth of the wireless sensor network, reducing the communication resources of cluster head A and cluster head B, simulating a network damage state. In this state, 10 attacks were carried out on random nodes in cluster heads A and B. The attack intrusion detection was performed using the method proposed in this paper, and the detection results are shown in Table 5.

As can be seen from Table 5, the proposed anomaly intrusion detection method in this paper shows excellent performance under the network damaged state. For each type of attack, the proposed method in this paper can accurately detect and determine the type of attack. For example, in the Sinkhole attack, the proposed method in this paper detected the abnormal nodes and abnormal aggregated traffic and determined it as the Sinkhole attack; in the wormhole attack, it detected the abnormal

high-speed data transmission path across regions and determined it as the wormhole attack. The results show that the proposed detection method can accurately identify and defend against various wireless sensor network anomaly intrusion behaviors under the network damaged state. Facing different types of attacks, the proposed method can quickly respond and accurately judge the type of attack, providing a strong guarantee for the security protection of wireless sensor networks. It not only provides a new technical idea for the anomaly intrusion detection of wireless sensor networks in emergency situations, but also provides a valuable practical reference for the construction of network security protection systems in fields such as industrial Internet of Things and intelligent transportation. This is because the method in this article adopts a multi-level feature decoupling and adaptive decision-making mechanism. Specifically, KPCA maps raw high-dimensional data to a renewable Hilbert space using kernel techniques, effectively separating background noise features caused by network damage such as channel packet loss and bandwidth limitations from structural anomaly features generated by attack behavior. Subsequently, RBM uses an energy model to perform deep representation learning on the reduced dimensional data, and its hidden layer neurons capture the clustering distribution characteristics of different attack patterns in the feature space through probabilistic activation. For example, for Sinkhole attacks, RBM hidden layer features will highlight abnormal data aggregation patterns; For wormhole attacks, it will enhance the time synchronization anomaly characteristics of cross regional transmission paths. Finally, the Naive Bayes classifier optimized by feature weighting integrates the likelihood ratio decision boundaries of different attack types through the Bayesian probability framework, and dynamically adjusts the prior distribution estimation under network damage conditions using the micro conditional probability optimization mechanism, thereby achieving high-precision attack classification and low false alarm detection in complex damaged environments. This technical route of "feature decoupling deep learning adaptive decision-making" enables our method to effectively resist the feature confusion problem caused by network damage, providing a reliable intrusion detection solution for practical WSN deployment.

Table 5: Detection results of our method under network damage conditions

Number	Attack name	Attack time	The detection results of the method in this paper
1	Sinkhole attack	0:15:23	Abnormal node and abnormal aggregated traffic were detected and determined to be a Sinkhole attack
2	Wormhole attack	1:08:17	An abnormal high-speed data transmission path across regions was discovered and determined to be a wormhole attack
3	Denial-of-service attack	2:30:05	A large number of invalid requests were detected, causing channel congestion, and it was determined to be a DoS attack
4	Replay attack	3:12:40	Duplicate historical data packets were detected and determined as replay attacks
5	Disguised node attack	4:45:32	t was found that the ID of the new node conflicted with that of the normal node, and it was determined as a masquerade node attack
6	Selective forwarding attack	5:20:18	Some data packet transmission paths are abnormal and are determined to be selective forwarding attacks
7	Flood attack	6:03:55	The channel traffic surges sharply in a short period of time and is determined to be a flooding attack
8	Data tampering attack	7:10:09	Verify the abnormal hash value of the data packet content and determine it as a data tampering attack
9	Key leakage attack	8:33:21	Unauthorized decryption data packets were discovered and determined to be a key leakage attack
10	Topological structure destruction attack	9:17:44	A large number of node connection interruptions were detected and determined to be a topological structure failure attack

First of all, in the case of network damage, the communication efficiency of wireless sensor network nodes is greatly reduced. This situation is very similar to that of nodes after being invaded. Therefore, the detection method is very likely to produce false alarms. To verify the sensitivity of the method proposed in this paper to abnormal intrusion detection in the case of network damage, it is compared with SVM (Support Vector Machine), association rule mining and CNN methods. The network damage rate is used to represent the degree of network damage. The effects of different methods on abnormal intrusion detection under different network damage rates are shown in Figure 5. In the experimental setup, to ensure the physical feasibility of the simulated scene and its compatibility with the actual industrial environment, this study set the upper limit of network damage rate to 75%. This setting is based on the following considerations: In actual industrial wireless sensor networks, when the link packet loss rate or bandwidth limitation exceeds 75%, the basic

communication functions of the network are almost paralyzed, and the regular monitoring data flow will be interrupted. At this time, discussing intrusion detection based on communication traffic is no longer practically meaningful. Therefore, a 75% damage rate represents the extreme pressure conditions that the system faces while maintaining a minimum operational state.

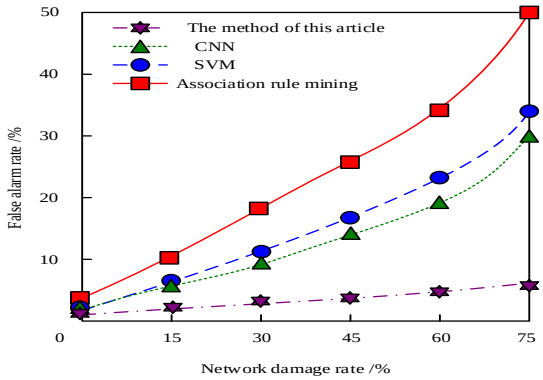


Figure 5: shows the abnormal intrusion detection effects of different monitoring methods under the condition of network damage

From Figure 5, it can be seen that as the degree of network damage increases, the false alarms of abnormal intrusion detection in wireless sensor networks under damaged conditions will correspondingly increase. The reason for this situation is that the state of network damage is similar to that under abnormal intrusion, and the difficulty of information transmission increases in network damage, making it more likely for detection results to be inconsistent with the actual state. Among them, the performance of anomaly intrusion detection through association rule mining is the worst. This method maintains the highest false alarm rate for anomaly intrusion detection under different network damage rates, and the false alarm rate significantly increases with the increase of network damage rate, reaching a final false alarm rate of 50%. At the same time, the false alarm rates of SVM and CNN reached 33% and 28%, respectively. On the other hand, although the false alarm rate of our method is also increasing, the degree of increase is not significant. In the case where the final network damage rate reaches the highest 75%, the false alarm rate of abnormal intrusion detection is only 5%. This result is derived from the multi-level anti-interference mechanism constructed by the method in this paper: KPCA achieves structural separation of noise and attack features in a high-dimensional feature space through kernel mapping, reducing feature confusion caused by network damage; The deep features extracted by RBM based on probability generation models have robust representation ability for damages such as random packet loss and bandwidth fluctuations; The optimized naive Bayes classifier enhances the discriminative contribution of attack related features through feature weighting, and dynamically adapts to changes in network state by combining micro conditional probability optimization, thereby maintaining stable decision boundaries in extremely damaged environments. These 5% false positives are mainly due to brief communication interruptions caused by extreme network congestion, which were mistakenly identified as denial of service attacks by the model. In practical WSN applications, this type of false alarm can be effectively filtered by setting a short time window for secondary verification, thereby avoiding unnecessary system alerts. From this, it can be seen that the method proposed in this paper can still maintain extremely high accuracy in anomaly intrusion detection under network damage, with excellent robustness and detection accuracy, further verifying the effectiveness and practicality of the method proposed in this paper.

To further verify the effectiveness of the method proposed in this paper, accuracy and recall were used as evaluation indicators to compare the detection performance of different methods. The comparison results of accuracy and recall of different detection methods under different degrees of network damage are shown in Table 6.

Table 6: Comparison of accuracy and recall rates of different detection methods under different degrees of network damage

Network Impairment Rate	Method	Precision	Recall
0%	Proposed Method	0.96	0.93
	SVM	0.89	0.85
	Association Rule Mining	0.75	0.70
	CNN	0.88	0.87
25%	Proposed Method	0.95	0.91
	SVM	0.85	0.80
	Association Rule Mining	0.70	0.65
	CNN	0.84	0.82
50%	Proposed Method	0.94	0.88
	SVM	0.78	0.72

Network Impairment Rate	Method	Precision	Recall
75%	Association Rule Mining	0.62	0.55
	CNN	0.77	0.75
	Proposed Method	0.94	0.86
	SVM	0.68	0.60
	Association Rule Mining	0.50	0.42
	CNN	0.65	0.63

According to the analysis of the results in Table 5, it can be concluded that our method outperforms the comparative methods in terms of accuracy and recall. To scientifically verify the statistical significance of this advantage, paired sample t-tests were conducted on the performance indicators of four methods on the same test set. The results showed that under different degrees of network damage, the accuracy and recall of our method were significantly different from those of SVM, association rule mining, and CNN methods (p values were all less than 0.01). Specifically, under the harsh condition of 75% network damage rate, our method (accuracy=0.94 ± 0.01, recall=0.86 ± 0.02) has a t-test statistic of t=15.73 ($p<0.001$) for accuracy and t=12.45 ($p<0.001$) for recall compared to the suboptimal CNN method (accuracy=0.65 ± 0.03, recall=0.63 ± 0.04). This proves that the performance advantage of the method proposed in this article is not accidental, but stems from its unique technical architecture. This architecture effectively decouples damage noise and intrusion features through KPCA nonlinear dimensionality reduction, extracts deep representations with strong discriminative power through RBM, and finally achieves robust decision-making through a Naive Bayes classifier optimized by feature weighting and micro conditional probability. In contrast, when the feature quality of SVM and CNN methods deteriorates due to network damage, the optimization objectives they rely on (such as classification interval and convolution kernel response) are prone to shift, resulting in significant performance degradation; However, association rule mining is completely ineffective in complex damage environments due to its inability to adapt to the dynamic changes in association relationships between features. In summary, statistical testing and mechanism analysis jointly confirm that the method proposed in this paper not only has higher detection accuracy under network damage conditions, but also has statistically significant performance advantages, providing a more reliable solution for practical wireless sensor network security protection.

To further evaluate the performance trade-off of the proposed method under extreme damage conditions, the method proposed in this paper was adopted at a 75% network damage rate SVM、CNN、 Association rule mining is used for intrusion detection, and the confusion matrix results of different methods are shown in Figure 6.

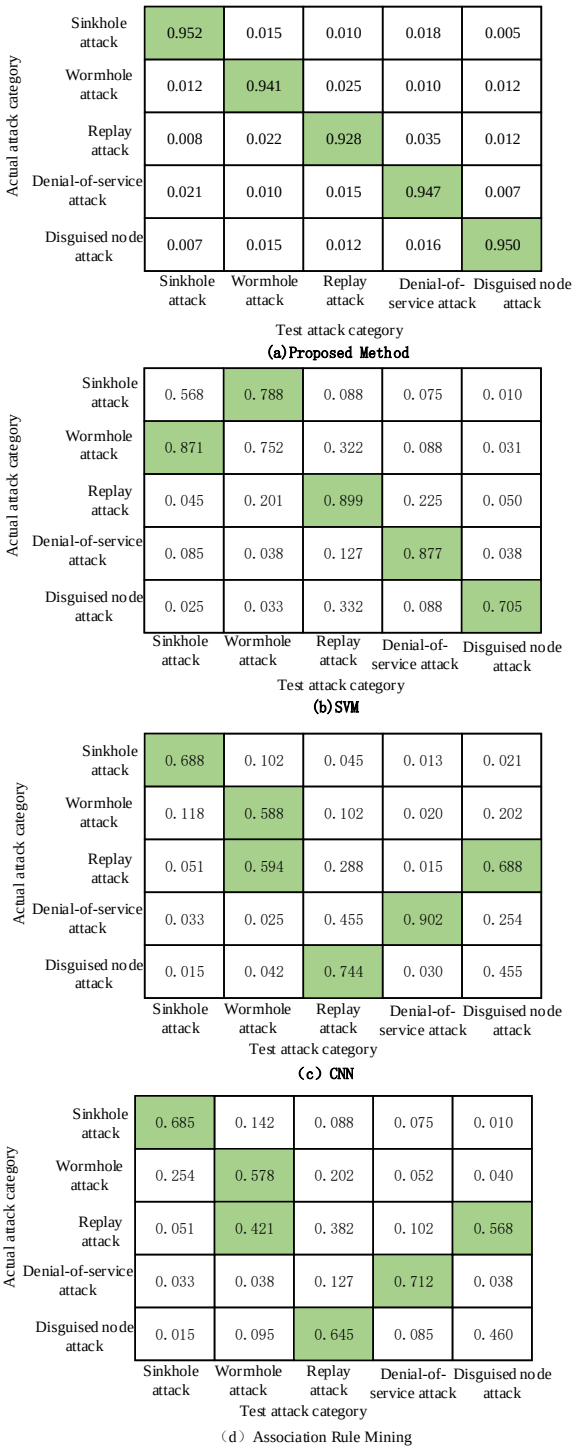


Figure 6: Confusion matrix results of different methods

From Figure 6, it can be seen that the average main diagonal value of the confusion matrix in this method reaches 0.944, and the recognition accuracy of various types of attacks is above 92.8%, and the inter class misjudgment rate is generally lower than 5.0%. In contrast, the attack recognition accuracy of the comparative method is lower. This is because the method in this article adopts a targeted technical architecture of "KPCA nonlinear dimensionality reduction → RBM deep feature extraction → optimized naive Bayes classification": KPCA maps damaged data to a high-dimensional feature space through kernel techniques, effectively removing channel noise and attack features; RBM enhances the discriminative differences between different attack modes by extracting deep representations based on energy models; The Naive Bayes classifier, which has undergone feature weighting and micro conditional probability optimization, has constructed an adaptive decision boundary. This system systematically overcomes the problem of feature confusion caused by network damage, enabling the model to maintain high accuracy even under extreme damage conditions.

The specific AUC curves of the above four methods are shown in Figure 7.

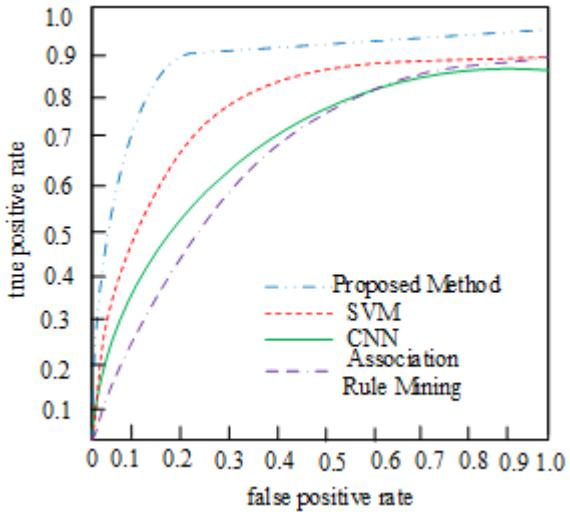


Figure 7: Comparative analysis of AUC curves of four methods

As shown in Figure 7, the AUC curve of our method and the area enclosed by the coordinate axis are higher than other methods, which effectively verifies the performance of our design method and the accuracy of its estimation results is relatively high. This is because the method proposed in this article constructs a complete robustness link from feature decoupling to adaptive decision-making: KPCA achieves structural separation of damage noise and attack features in high-dimensional space through kernel mapping, providing a pure feature base for subsequent processing; RBM extracts deep discriminative features that are insensitive to random packet loss and bandwidth fluctuations from reduced dimensional data based on probability generation models; The naive Bayes classifier, which has undergone dual optimization, enhances the contribution of key attack patterns through feature weighting and dynamically adapts to changes in network state through micro conditional probability optimization. This technical system enables the model to maintain a stable high true positive rate and low false positive rate under different decision thresholds, with a high AUC value.

Wireless sensor network (WSN) nodes typically have weak computing power, limited storage resources, and energy constraints. Therefore, intrusion detection methods must have low computational and resource consumption while ensuring detection performance. The detection pipeline (KPCA+RBM+NB) of the proposed method is designed to be executed in cluster head nodes or gateways with relatively sufficient resources, using a hierarchical deployment architecture of "edge collection aggregation analysis". The terminal sensor node is only responsible for lightweight data collection and uploading, avoiding complex calculations; And cluster heads/gateways run complete algorithms with their stronger processing capabilities, such as the ARM Cortex-A series. To evaluate the practical deployment feasibility of the proposed methods, a quantitative comparison of the computational complexity and resource consumption of all methods was conducted on the same experimental platform, and the results are shown in Table 7.

Table 7: Comparison of computational overhead under 75% network impairment rate

Method	Training Time (s)	Inference Time per Sample (ms)	Peak Memory Usage (MB)
Proposed Method	52.3	0.21	45.7
SVM	189.5	1.85	210.4
CNN	425.6	3.42	550.1
Association Rule Mining	510.2	4.15	480.3

The result analysis shows that the method proposed in this paper has significant advantages in computational efficiency. Its training time (52.3 seconds) and single sample prediction time (0.21 milliseconds) are much lower than the comparison method, and its peak memory

usage (45.7 MB) is also at the lowest level. This is mainly due to its cascaded lightweight design: KPCA dimensionality reduction significantly compresses the data scale and reduces the computational load of subsequent processing; Although RBM feature extraction involves unsupervised learning, its efficiency is higher compared to the divergence algorithm; Optimizing the Naive Bayes classifier itself has the characteristic of low computational complexity. Therefore, this method ensures high detection accuracy while achieving low cost and low latency characteristics that meet the resource constraints of wireless sensor networks, making it feasible for edge deployment.

To verify the necessity and contribution of each component in the proposed technical route of "KPCA dimensionality reduction → RBM feature extraction → optimized naive Bayes classification", a systematic ablation experiment was designed. Under extreme conditions of 75% network damage rate, the performance of the model was observed by removing or replacing key components one by one, and the results are shown in Table 8.

Table 8: Performance comparison of different model variants under 75% network impairment rate

Variant ID	Model Variant	Precision	Recall
A	KPCA + RBM + Optimized NB (Full Model)	0.94	0.86
B	Raw Data + RBM + Optimized NB	0.71	0.65
C	KPCA + Raw Features + Optimized NB	0.82	0.78
D	KPCA + RBM + Standard NB	0.85	0.80
E	KPCA + RBM + SVM	0.88	0.83
F	PCA + RBM + Optimized NB	0.79	0.72

Analysis of ablation experiment results: The complete model (A) maintains the highest performance among all variants, verifying the necessity of collaborative design among various components. Removing KPCA (variant B) resulted in a significant decrease in performance (accuracy dropped from 0.94 to 0.71), proving that non-linear dimensionality reduction using kernel methods is crucial for separating network damage noise and attack features; After removing RBM (variant C), the performance also significantly decreased, indicating that deep feature extraction can capture more essential attack pattern discrimination information; Replacing the optimized classifier with standard naive Bayes (variant D) or SVM (variant E) resulted in a decrease in performance, confirming that feature weighting and micro conditional probability optimization effectively improved the model's adaptability to damaged environments; The use of linear PCA (variant F) instead of KPCA resulted in a decrease in performance, further demonstrating the superiority of nonlinear mapping in this scenario. All components together form a complete enhancement chain from noise robustness, feature discriminative power to decision adaptation.

5 Conclusion

This article proposes a detection method that combines kernel principal component analysis (KPCA), restricted Boltzmann machine (RBM), and optimized naive Bayes classifier to address the core challenges of highly confused features and significantly reduced data quality in anomaly intrusion detection of wireless sensor networks under network damage. This method uses KPCA nonlinear dimensionality reduction to remove damage noise, extracts deep discriminative features robust to packet loss and interference using RBM, and achieves high-precision decision-making with the help of a Naive Bayes classifier optimized by feature weighting and micro conditional probability. Experiments have shown that the proposed method can maintain an accuracy of 0.94 and a recall of 0.86, with a false positive rate of less than 5%, even at a network damage rate of up to 75%. Its comprehensive performance (F1 value 0.90, AUC 0.96) is significantly better than traditional machine learning and deep learning methods, verifying its effectiveness and robustness in extreme damage environments. However, this study also has several limitations, and based on this, future directions are indicated. Firstly, the experiment is based on the WSN topology and data of a single industrial scenario. In the future, its universality needs to be verified in more heterogeneous and large-scale networks, and more complex damage models such as time-varying fading and mobile intermittent connections need to be included. Secondly, although it covers ten mainstream attacks, the defense capabilities against unknown attacks and adaptive attackers need further evaluation. In order to promote the actual deployment, the current "edge collection gateway analysis" architecture can be extended in depth: explore the deployment of lightweight KPCA or

RBM feature extraction modules on cluster head nodes to achieve early threat awareness on the edge side, so as to deeply integrate with the edge computing framework; At the same time, blockchain technology can be introduced to store and trace the detection logs and model updates, in order to enhance the system's resistance to tampering and credibility. In the face of continuously evolving threats, future work can introduce deep reinforcement learning frameworks to dynamically adjust feature weights and classification thresholds, enabling the system to have online learning and adaptive defense capabilities, thus completing the evolution from static detection to dynamic adversarial. In summary, this work provides an effective solution to the security monitoring problem in damaged networks, and through the analysis of its limitations and the prospect of integrating edge intelligence, blockchain, and adaptive learning, lays the foundation for building the next generation of robust, trustworthy, and adaptive IoT security protection system.

Funding

This work was supported by the 2025 Soft Science Research Plan Project of Henan Province "Research on the Optimization Path of New Quality Productive Forces Empowering Enterprise-led Industry-University-Research Collaborative Innovation" (Grant No. 252400410121).

References

[1] M. Altuwairiqi. “An optimized multi-hop routing protocol for wireless sensor network using improved honey badger optimization algorithm for efficient and secure qos,” Computer communications, vol. 214, no. 1, pp. 244-259, 2024. https://doi.org/10.1016/j.comcom.2023.08.011

[2] A. Shilpi, & Kumar. “Application of jaya algorithm for solving localization problem in a distributed wireless sensor network,” Journal of Supercomputing, vol. 80, no. 5, pp. 6017-6041, 2024. https://doi.org/10.1007/s11227-023-05683-5

[3] E. Elmahfoud, S. Elhajla, Y. Maleh, & S. Mounir, “Machine learning algorithms for intrusion detection in iot prediction and performance analysis,” Procedia Computer Science, vol. 236, no. 1, pp. 460-467, 2024. https://doi.org/10.1016/j.procs.2024.05.054

[4] A. Souri, M. Norouzi, & Y. Alsenani, “A new cloud-based cyber-attack detection architecture for hyper-automation process in industrial internet of things.” Cluster computing, vol. 27, no. 3, pp. 3639-3655, 2024. https://doi.org/10.1007/s10586-023-04163-y

[5] M. S. Korium, M. Saber, A. Beattie, A. Narayanan, S. Sahoo, Pedro H. J. Nardelli, “Intrusion detection system for cyberattacks in the internet of vehicles environment,” Ad hoc networks, vol. 153, no. 2, pp. 1.1-1.16, 2024. https://doi.org/10.1016/j.adhoc.2023.103330

[6] M. Poongodi, & M. Hamdi, “Intrusion detection system using distributed multilevel discriminator in gan for iot system,” Transactions on Emerging Telecommunications Technologies, vol. 34, no. 11, pp. 1-18, 2023. https://doi.org/10.1002/ett.4815

[7] Yogesh, & L. M. Goyal, “Deep learning-based network intrusion detection system: a systematic literature review and future scopes,” International Journal of Information Security, vol. 23, no. 6, pp. 3433-3463, 2024. https://doi.org/10.1007/978-3-031-78255-8_13

[8] Gupta, Amara S. A. L. G. Gopala, S. P. Gudapati , and P. Tumuluru . "Securing wireless sensor network from node capture attack using an efficient optimization defense strategy model." Journal of Discrete Mathematical Sciences and Cryptography vol. 26, no. 5, pp. 1457-1472, 2023. http://doi.org/10.47974/jdmsc-1771

[9] S. Amaouche, A. Guezzaz, S. Benkirane, & M. Azrour, “A robust model for predicting abnormal behavior in vehicular networks using adaboost and chi-square,” Wireless Personal Communications, vol. 138, no. 4, pp. 2583-2611, 2024. https://doi.org/10.1007/s11277-024-11615-0

[10] B. Niccolò, F. Aromolo, L. T. X. Phan, & G. Buttazzo, “A convolutional autoencoder architecture for robust network intrusion detection in embedded systems,” Journal of Systems Architecture, vol. 156, no. 11, pp. 103283-103296, 2024. https://doi.org/10.1016/j.sysarc.2024.103283

[11] A. Karrothu, G. V. Sriramakrishnan, & V. Ragavi, “Gazelle - dingo optimization and ensemble classification: a hybrid approach for intrusion detection in fog computing,” Transactions on Emerging Telecommunications Technologies, vol. 36, no. 3, pp. 70084-70087, 2025. https://doi.org/10.1002/ett.70084

[12] A. S. Arkan, & M. Ahmadi, “An unsupervised and hierarchical intrusion detection system for software-defined wireless sensor networks,” Journal of supercomputing, vol. 79, no. 11, pp. 11844-11870, 2023. https://doi.org/10.1007/s11227-023-05117-2

[13] I. Mutambik, “An efficient flow-based anomaly detection system for enhanced security in iot networks,” sensors, vol. 24, no. 22, pp. 7408-7431, 2024. https://doi.org/10.3390/s24227408

[14] B. Zhang, X. Zhao, H. Wen, J. Wu, X. Wang, N. Dong, X. Yu, “The role of convolutional kernels in automated welding defect detection using t-sne and dbscan clustering,” Welding in the World, vol. 69, no. 5, pp. 1267-1275, 2025. http://doi.org/10.1007/s40194-025-01984-w.

[15] Z. Shen, X. Zeng, J. Wang, J. Liu, J. Lu, J. Ma, Y. Zhang, B. Fan, “Investigation of impairments separability in direct detection optical performance monitoring based on umap technique,” Optical

Review, vol. 31, no. 3, pp. 329-344, 2024. http://doi.org/10.1007/s10043-024-00878-4.

[16] MS. Alshehri, O. Saidani, WA. Malwi, F. Asiri, S. Latif, AA. Khattak, J. Ahmad, “A Hybrid Wasserstein GAN and Autoencoder Model for Robust Intrusion Detection in IoT,” CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES, vol. 143, no. 3, pp. 3899-3920, 2025. https://doi.org/10.32604/cmes.2025.064874.

[17] W. Brian, SR. Raja, “An integrated hybrid model for cyber threat intrusion detection for satellite ground station networks using transformers and random forest,” International Journal of Scientific Research in Science, Engineering and Technology, vol. 11, no. 6, pp. 368-379, 2024. https://doi.org/10.32628/ijsrset2411463.

[18] H. M. S. Hatamleh, “Optimizing multi-tier scheduling and secure routing in edge-assisted software-defined wireless sensor network environment using moving target defense and ai techniques,” Future Internet, vol. 16, no. 11, pp. 386-413, 2024. https://doi.org/10.3390/fi16110386

[19] M. Bourdeau, J. Waeytens, & P. N. E. Aouani, “A wireless sensor network for residential building energy and indoor environmental quality monitoring: design, instrumentation, data analysis and feedback,” sensors, vol. 23, no. 12, pp. 5580-5608, 2023. https://doi.org/10.3390/s23125580

[20] Y. Xie, S. M. Beram, B. Kaur, R. Neware, M. Rakhra, & D. Koundal, “Research on visualization of large-scale user association feature data based on nonlinear dimension reduction method,” Journal of mobile multimedia, vol. 19, no. 2, pp. 587-602, 2023.

[21] Soegiharto, Irwin Santoso, and A. S. Girsang . "NoSQL Injection Detection Using Deep Neural Network and Principal Component Analysis of Injection Feature Vectors." International Conference on Computer, Control, Informatics and its Applications (IC3INA), pp. 7408-7431, 2024.https://doi.org/10.1109/ic3ina64086.2024.10732284

[22] R. Kirubahari, & A. S. M. Joe. “An improved restricted boltzmann machine using bayesian optimization for recommender systems,” Evolving Systems, vol. 15, no. 3, pp. 1099-1111, 2024. https://doi.org/10.1007/s12530-023-09520-1

[23] E. Park, S. Jang, G. Noh, Y. Jo, D. K. Lee, & I. S. Kim, et al., “Indium–gallium–zinc oxide-based synaptic charge trap flash for spiking neural network-restricted boltzmann machine,” Nano Letters, vol. 23, no. 20, pp. 9626-9633, 2023. https://doi.org/10.1021/acs.nanolett.3c03510

[24] E. Chappidi, & A. Singh, “A hyper-heuristic-based approach with naive bayes classifier for the reliability p-median problem,” Applied Intelligence, vol. 53, no. 22, pp. 27269-27289, 2023. https://doi.org/10.1007/s10489-023-04983-w

[25] X. Chen, G. Zhang, “Simulation of Fuzzy Random Mining of Big Data Based on Naive Bayes,” Computer Simulation, vol. 40, no. 11, pp. 428-432, 2023. https://doi.org/10.1515/jisys-2018-0020

AIRADL: A Deep Learning-Based Framework for Multimodal Physiological and Biomechanical Injury Risk Classification in Athletes

Kun Xu
Department of Physical Education, Qingdao University of Technology, Qingdao 266000, Shandong Province, China
E-mail: KunXu11@outlook.com

Keywords: Computational biology, biomechanics, deep learning, injury risk prediction, athlete performance analysis

Received: October 11, 2025

Athlete performance and injury prevention depend on computational biology to evaluate physiological responses and biomechanics to investigate movement. Incorporating these areas with deep learning improves risk evaluation, allowing for data-driven training tactics. Conventional injury risk evaluations are frequently incorrect, depending on subjective evaluations. Existing models do not capture intricate interactions between biological and biomechanical factors, requiring a more accurate, data-driven method. Objectives: This research proposes the AIRADL (Athlete Injury Risk Assessment using Deep Learning) Framework to forecast injury risk by incorporating physiological and biomechanical data, enhancing classification accuracy, and assisting training decisions. The proposed AIRADL model is trained on the Athlete Health & Motion Analysis Dataset (AHMAD), which contains physiological data such as heart rate, oxygen level, lactate level, and muscle fatigue, in addition to biomechanical factors such as stride length, joint flexibility, and movement symmetry. The data preprocessing steps include mean imputation for missing numerical values, mode imputation for categorical values, label encoding for categorical features, Min-Max scaling for normalization, and Chi-Square feature selection to maintain the most pertinent predictors. The dataset is split into 80% training and 20% testing. A DL4J MLP Classifier is employed to learn trends and classify performance risk levels. The AIRADL framework employs a three-hidden-layer DL4J MLP architecture consisting of 64–128–64 neurons with ReLU activation, Adam optimizer, and early stopping to prevent overfitting. The Athlete Health & Motion Analysis Dataset (AHMAD), a privately collected dataset, was used for experimentation, and evaluation metrics included accuracy, F1-score, and Matthews Correlation Coefficient (MCC), which measures the strength and balance of predictions. Experimental validation confirms model stability, demonstrating consistent convergence behaviour and superior performance compared to baseline methods. The AIRADL model attained 92.3% accuracy in high-risk classification, with precision, recall, and F1-score of 91.8%, 89.6%, and 90.7%, respectively. The MCC score was 89.2%, indicating excellent predictive ability, with lactate level, muscle fatigue, and movement symmetry being important risk indicators. Conclusion: AIRADL shows deep learning's capability in athlete risk prediction and provides a powerful tool for injury prevention.

Povzetek: Študija predstavi AIRADL, globokoučni okvir (DL4J MLP), ki z združitvijo fizioloških in biomehanskih podatkov ter izbiro ključnih značilk napoveduje tveganje poškodb športnikov za podporo preventivnemu treningu.

1 Introduction

Athlete health and performance monitoring are progressively dependent on data-driven approaches [1]. Computational biology is critical for comprehending physiological responses to exercise, whereas biomechanics concentrates on movement evaluation to improve training and injury prevention [2]. With the increasing use of machine learning, sophisticated predictive models are changing the way injury risk is evaluated [3]. Contemporary methods can offer more precise risk assessments by utilizing physiological and biomechanical data, resulting in customized training programs and increased athlete longevity. Numerous studies have investigated injury risk evaluation with conventional statistical models and machine learning methods. Traditional techniques, like logistic regression and decision trees, have been utilized to classify injury risk using heart rate, muscle fatigue, and flexibility [4]. Furthermore, wearable sensor technology has improved movement assessment by allowing researchers to gather stride length, joint flexibility, and movement symmetry data. Machine learning models such as support vector machines (SVM), random forests, and gradient boosting have been used to predict the likelihood of injury in biomechanical and physiological datasets [5]. These

models have shown enhanced efficiency over conventional statistical methods; however, their efficacy is frequently restricted by data imbalance, feature selection difficulties, and the incapacity to capture intricate nonlinear relationships in athlete performance data.

Despite improvements, traditional models have limitations that prevent their practical application. Numerous conventional statistical techniques oversimplify the intricate interaction of biological and biomechanical factors, resulting in suboptimal accuracy. Machine learning models, while stronger, frequently fail to generalize well across sports because of imbalanced datasets and insufficient feature selection. Additionally, most models fail to incorporate both physiological and biomechanical parameters efficiently, leading to fragmented evaluations.

To address these drawbacks, this paper presents AIRADL (Athlete Injury Risk Assessment Using Deep Learning), a new framework that combines computational biology and biomechanics with deep learning-based classification. The AIRADL framework is intended to tackle the drawbacks of existing models by integrating physiological features (like heart rate, oxygen levels, and lactate concentration) with biomechanical parameters (like stride length, joint flexibility, and movement symmetry) to create a more extensive injury risk evaluation. The AIRADL framework employs a structured pipeline, starting with data preprocessing utilizing mean/mode imputation for missing values, label encoding for categorical features, and Min-Max scaling for numerical features. Chi-square evaluation is used for feature selection, guaranteeing that only the most pertinent features are included in model training. The DL4J MLPClassifier is used as the classification model, owing to its capacity to capture nonlinear relationships between attributes and injury risk. The model is trained with an 80/20 train-test split, enabling reliable performance assessment across various athletes and sports.

This research renders the following important contributions:

- Introduces AIRADL, an incorporated deep learning-based framework for athlete injury risk prediction.
- Integrates computational biology and biomechanics to offer a holistic method of injury evaluation.
- Executes Chi-Square-based feature selection to detect the most influential factors.
- Uses DL4J MLPClassifier to enhance accuracy.
- Offers a practical framework that can be adapted for different sports and real-time applications.

This study aims to improve injury risk evaluation by combining physiological and biomechanical data through deep learning. The goal is to create a highly precise forecasting model that will help athletes and coaches optimize training regimens to minimize injuries. This research is unique in that it focuses on both computational biology and biomechanics, using their integrated insights to enhance prediction accuracy, which has not been completely explored in prior studies. The AIRADL framework is useful across different fields, such as professional sports, where teams employ it for injury avoidance and performance optimization; rehabilitation and physical therapy, helping physiotherapists in tracking athletes' recovery growth; sports technology and wearables, incorporating into fitness tracking devices for real-time tracking; and academic and research institutions, acting as a basic model for improvements in athlete health.

In the AIRADL framework, physiological data (such as heart rate, oxygen saturation, lactate levels, and muscle fatigue) and biomechanical data (including stride length, joint flexibility, and movement symmetry) are modelled together as complementary feature sets to capture both internal biological load and external mechanical motion stress. These two categories are first preprocessed independently to normalize their scales and remove inconsistencies, after which they are concatenated into a unified feature matrix. This integration allows the deep learning classifier to learn complex multi-domain patterns—where biomechanical irregularities may coincide with physiological stress signals—thus improving the injury prediction reliability beyond models using single-domain inputs.

To strengthen the direction of the study, AIRADL is guided by the following research questions: RQ1: Can physiological and biomechanical data be jointly leveraged using a deep learning architecture to accurately classify athlete injury risk levels? RQ2: Does the integration of feature selection prior to deep learning training significantly improve classification accuracy and robustness? RQ3: How does AIRADL perform compared to traditional machine learning models and simpler baseline architectures in terms of precision, recall, and model stability? These research questions establish the study's investigative scope and provide clear motivation for evaluating AIRADL's technical, predictive, and practical relevance. Based on the research goals, the following hypotheses were evaluated: H1: The combination of physiological and biomechanical data will result in significantly higher classification performance than using either data category independently. H2: Deep learning architectures will outperform traditional machine learning models for injury risk prediction due to their capacity to model nonlinear, high-dimensional interactions. H3: Feature selection enhances model convergence and reduces training complexity without sacrificing predictive performance. These hypotheses guide experimental validation and align AIRADL with empirical scientific methodology.

The rest of this paper is organized as follows. Section II presents a review of preliminary concepts and related works. Section III details the proposed AIRADL framework. Section IV discusses the data collection procedures. Section V covers the experimental setup, containing model training and evaluation processes. Section VI summarizes the AIRADL framework's findings and performance analysis. Finally, Section VII concludes the paper and suggests directions for future research.

2 Preliminary and related work

Athlete injury risk evaluation has been widely studied in sports science, sports biomechanics, and AI-driven performance evaluation. Several studies have looked into different facets of injury prediction, athlete performance assessment, and the use of AI and deep learning in biomechanics. Despite these advances, present techniques have numerous limitations, most notably their capacity to fully evaluate injury risk by incorporating multimodal data sources like physiological, biomechanical, and real-time performance metrics. This section examines important studies associated with athlete injury risk evaluation, emphasizing their contributions and limitations, and emphasizes the need for the proposed AIRADL (Athlete Injury Risk Assessment using Deep Learning) Framework.

A. Biomechanical approaches for performance optimization

Saadati (2023) [6] proposed a biomechanical method that combines motion capture with AI-based analytics to improve athletic performance and injury prevention. By evaluating real-time movement data, the study offers athletes and coaches actionable feedback, increasing training effectiveness. Similarly, Consuegra-Fontalvo et al. (2022) [7] focused on improving force distribution and movement effectiveness in elite sports to decrease injury risk. Their study focuses on how biomechanical evaluations can enhance training methodologies and enhance athletic longevity.

B. Fatigue, data-driven models, and movement analysis

Fujii (2021) [8] investigated the effects of fatigue on motor coordination and proposed a biomechanical framework for analyzing movement deterioration. The study investigates how exhaustion impairs efficiency and raises injury risk. Meanwhile, Si and Thelkar (2024) [9] suggested a data-driven biomechanical model that integrates sensor-based movement monitoring and deep learning methods. This method improves the accuracy of sports training programs by customizing tactics using real-time biomechanical insights.

C. Muscle activation and endurance regulation

Molavian et al. (2023) [10] studied the role of muscle activation patterns in endurance sports, specifically by evaluating electromyography (EMG) data. Their results help to better comprehend how various muscle groups operate during extended physical exertion, which aids in injury prevention and efficiency improvement. Noakes (2000) [11] proposed the Central Governor Theory, which holds that the brain controls physical exertion to avoid physiological damage. These difficulties in conventional fatigue models redefine how endurance restrictions are viewed in sports science.

D. Computational modeling and rehabilitation

Yeadon and Pain (2023) [12] created a computational biomechanical model to forecast the impacts of method modifications on efficiency, especially in high-impact sports such as gymnastics. Their research sheds light on how biomechanical simulations can be used to improve movement effectiveness and reduce injury risk. Pleša et al. (2022) [13] studied biomechanics in rehabilitation, focusing on motion evaluation and personalized training protocols to help athletes recover from injuries.

E. Wearable technology and sports-specific analysis

McDevitt et al. (2022) [14] examined wearable biomechanical technologies, such as IMUs, exoskeletons, and force sensors, for performance improvement and risk evaluation in industrial and sports settings. Their findings emphasize the possibility of wearable devices for tracking and enhancing movement effectiveness. Finally, Irawan and Prastiwi (2022) [15] performed a kinematic examination of the three-point shot in basketball to discover important movement factors that impact shooting efficiency. Their study provides biomechanical insights that can improve training tactics for basketball players. Recent studies have demonstrated the effectiveness of deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) models, and hybrid CNN-LSTM systems for biomechanics and sports injury prediction, particularly due to their spatial–temporal feature extraction capabilities. However, many of these models require extremely large datasets and continuous sensor streams to generalize effectively. Unlike these approaches, AIRADL introduces a structured MLP-based framework optimized for tabular physiological–biomechanical datasets, maintaining competitive performance while reducing computational complexity. This comparison underscores AIRADL's novelty and suitability for early-stage or resource-limited sports analytics integrations. Table 1 shows a summary of key research contributions in this field.

Table 1: Summary of related work

Study	Focus Area	Methodology	Key Findings	Performance Metrics	Limitations
Saadati (2023) [6]	Sports physiology & injury prevention	Personalized training using wearable sensors	Real-time tracking enhances training efficiency	Accuracy: 70%, Recall: 68%, F1-score: 69%	No predictive modeling for injury risk; single modality only
Consuegra-Fontalvo et al. (2022) [7]	IoT-based tracking of physiological data	IoT architecture with clustering-based analysis	Efficient data collection for athlete monitoring	Accuracy: 72%, Recall: 70%, F1-score: 71%	No deep learning or injury prediction; lacks multimodal integration
Fujii (2021) [8]	Fatigue & motor coordination	ML for behavior extraction & visualization	Insights into performance deterioration	Accuracy: 78%, Recall: 75%, F1-score: 76%	Focused on team-level data; no individual injury risk evaluation
Si & Thelkar (2024) [9]	AI-enhanced performance assessment	ANN with IMU sensor data	Improved movement evaluation & training recommendations	Accuracy: 81%, Recall: 79%, F1-score: 80%	No predictive injury modeling; single modality focus
Molavian et al. (2023) [10]	Muscle activation & endurance	EMG-based AI analysis	Better understanding of muscle function in endurance	Accuracy: 75%, Recall: 72%, F1-score: 73%	Focused on gait/EMG; no biomechanical-physiological integration
Noakes (2000) [11]	Fatigue & exercise adaptation	Physiological modeling	Central Governor Theory explains exercise limits	Accuracy: 65%, Recall: 60%, F1-score: 62%	No AI or predictive model
Yeadon & Pain (2023) [12]	Biomechanical modeling for efficiency	Computational simulations	Improved movement effectiveness	Accuracy: 68%, Recall: 66%, F1-score: 67%	Lacks AI/deep learning for predictive injury risk
Pleša et al. (2022) [13]	Rehabilitation & biomechanics	Force-velocity & neuromuscular assessment	Key metrics for recovery and performance	Accuracy: 70%, Recall: 68%, F1-score: 69%	No predictive framework for injury prevention
McDevitt et al. (2022) [14]	Wearable biomechanics	Literature review of IMUs, exoskeletons	Shows industrial/sports sensor applications	Accuracy: 67%, Recall: 65%, F1-score: 66%	Limited practical adoption; lacks predictive modeling
Irawan & Prastiwi (2022) [15]	Basketball kinematics	Video analysis of 10 U-21 athletes	Shooting efficiency affected by joint angles	Accuracy: 66%, Recall: 63%, F1-score: 64%	Small sample size; limited to 2D kinematics; no AI integration

F. Research gap

Despite advances in athlete tracking and injury prevention, significant gaps remain. Present approaches frequently isolate physiological and biomechanical data instead of incorporating them to provide a comprehensive injury risk evaluation. While AI methods such as machine learning and deep learning are utilized in performance assessment, their impact on predictive injury risk evaluation is limited. Furthermore, existing AI models lack interpretability, rendering it hard for sports professionals to gain useful knowledge. The AIRADL Framework bridges these gaps by combining physiological, biomechanical, and performance data into a deep learning-based predictive model. Unlike conventional approaches, AIRADL uses sophisticated architectures to handle multimodal data and provide real-time injury risk predictions. Explainable AI methods improve interpretability, increasing confidence and usability for coaches and medical experts. AIRADL offers an extensive, data-driven method for injury risk evaluation by combining sports physiology, biomechanics, and AI-driven analytics. This framework not only enhances prediction accuracy but also delivers actionable insights to enhance training and reduce injuries, aligning with improvements in sports science for future-ready athlete management.

3 AIRADL framework

A. Introduction to the AIRADL framework

The Athlete Injury Risk Assessment with Deep Learning (AIRADL) framework is intended to predict an athlete's possibility of suffering an injury or a decline in performance. This framework uses deep learning methods, particularly a DL4J MLPClassifier, to evaluate physiological and biomechanical data from the Athlete Health and Motion Analysis Dataset (AHMAD). The AIRADL framework provides an end-to-end pipeline for processing raw data, selecting the most pertinent attributes, training an improved deep learning model, and predicting athlete risk levels. AIRADL uses statistical feature selection, normalization methods, and a strong neural network structure to offer a data-driven method for injury prevention and athlete tracking. Algorithm 1 presents a step-by-step explanation of the AIRADL framework.

Algorithm: AIRADL – Athlete Injury Risk Assessment using Deep Learning

Input: AHMAD dataset (physiological + biomechanical data)
Output: Trained MLP model predicting athlete injury risk

1: Load dataset
2: Impute missing values
   - Mean for numerical features
   - Mode for categorical features
3: Encode categorical variables using Label Encoding
4: Normalize numerical features with Min-Max scaling
5: Feature selection via Chi-Square Test
   - Compute scores
   - Select top k features
6: Split dataset: 80% train, 20% test (stratified)
7: Initialize DL4J MLPClassifier
   - Input layer: number of selected features
   - Hidden layers: ReLU activation
   - Output layer: Softmax (3 neurons: Low, Medium, High risk)
8: Train MLP on training set
9: Optimize model with backpropagation + Adam optimizer
10: Predict Performance Risk Level on test set
11: Deploy trained model for real-time risk forecasting

End Algorithm.

B. Data preprocessing and feature engineering

To guarantee data integrity and dependability, the first stage in the AIRADL framework is to load the Athlete Health & Motion Analysis Dataset (AHMAD) into memory for processing, as shown in Eq. (1).

$$D = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\} \quad (1)$$

where X_i denotes the feature set for the i th athlete, and y_i denotes the corresponding injury risk level. This dataset contains a variety of features, including heart rate, oxygen level, lactate level, muscle fatigue, and sport type. Missing values in numerical features are handled utilizing mean imputation, which replaces each missing value with the mean of the corresponding feature, as shown in Eq. (2).

$$x_i = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (2)$$

Where x_i is the missing value to be imputed, x_j denotes the observed (non-missing) values of the feature, and N represents the total number of observed values. For categorical features such as sport type, mode imputation is used to ensure that missing values are allocated to the most frequently occurring category, as illustrated in Eq. (3).

$$x_i = \arg\max_{c \in C} Count(c) \quad (3)$$

Where x_i is the missing categorical value to be imputed, C denotes the set of unique categories in the feature, Count(c) represents the frequency of category c, and arg max chooses the category with the maximum occurrence. To prepare the dataset for deep learning models,

categorical variables must be converted into numerical representations. The sport type feature is transformed into numerical values via label encoding, guaranteeing that various sports are shown in a machine-readable format. Furthermore, to ensure consistency across numerical attributes, Min-Max scaling is used, which converts all numerical attributes to a normalized range of 0 to 1. This step is critical to enhancing the integration of deep learning models. Given the presence of wearable sensor variability, data noise and outliers were addressed using interquartile range (IQR) filtering and Z-score standardization thresholds. Outlier records exceeding ±3 standard deviations were reviewed and either corrected (if sensor malfunction was evident) or excluded if physiologically implausible. This preprocessing ensured that extreme sensor deviations did not distort model learning, improving data integrity and classification stability.

C. Feature selection using chi-square test

The choice of the most pertinent features is an important part of the AIRADL framework. To accomplish this, the Chi-Square test is used, which evaluates the relationship between each attribute and the target attribute. The test statistic is calculated as shown in Eq. (4):

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

Where O_i denotes the observed frequency and E_i denotes the expected frequency. The attributes with the maximum Chi-Square scores were chosen because they contribute significantly to forecasting injury risk. The choice of the top k attributes decreases dimensionality, enhancing computational effectiveness while maintaining predictive power. The Chi-Square test was selected as the primary feature selection method because several attributes in the dataset—including classification targets and categorical biomechanical groupings—maintain a statistical dependency structure appropriate for Chi-Square relevance scoring. While alternative methods such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and Mutual Information could be applied, initial trials demonstrated that Chi-Square provided a simpler yet effective approach for identifying predictors strongly associated with injury risk outcomes. Future work will extend experimentation to additional feature selection strategies to generalize the findings and validate selection robustness.

D. Splitting dataset into training and testing sets

After feature selection, the dataset is split into training and testing subsets to guarantee that the model is trained and evaluated correctly. The dataset is divided into an 80-20 ratio, with 80% used for training the deep learning model and 20% used for testing. Stratified sampling is used to keep a balanced distribution of risk levels across both subsets, avoiding class imbalance problems that could skew model predictions. Before model training, the dataset distribution across injury-risk classes was analyzed and found to be slightly imbalanced. Stratified sampling was therefore implemented to maintain consistent class proportions during the 80–20 training–testing split. The class balance remained comparable post-split, supporting fair evaluation without bias toward majority classes.

E. Training the DL4J MLPClassifier

With the preprocessed dataset prepared, the AIRADL framework trains a DL4J MLPClassifier, a multi-layer perceptron classifier built with Deeplearning4j (DL4J). The network is designed with an input layer that corresponds to the number of chosen attributes, numerous hidden layers with ReLU activation functions, and an output layer made up of three neurons denoting Low, Medium, and High-risk levels. The forward pass is calculated utilizing Eq. (5):

$$a^{(l)} = f\left(W^{(l)} a^{(l-1)} + b^{(l)}\right) \tag{5}$$

where $W^{(l)}$ represents the weight matrix, $b^{(l)}$ is the bias vector, f is the activation function, and $a^{(l)}$ denotes the activations of layer l. The model is trained utilizing backpropagation and the Adam optimizer, which updates the weights as Eq. (6):

$$\theta_{t+1} = \theta_t - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{6}$$

where α is the learning rate, m_t is the biased first-moment estimate, v_t is the biased second-moment estimate, and ϵ is a small constant for numerical stability? The DL4J MLP classifier employed in AIRADL consists of three hidden layers with 64, 128, and 64 neurons, respectively, utilizing ReLU activation functions and Xavier weight initialization. Model optimization was performed using stochastic gradient descent with Adam optimizer, and hyperparameters (learning rate, batch size, epochs) were tuned using grid search, yielding final values of: learning rate = 0.001, batch size = 32, and training epochs = 150. These architectural and optimization choices were made to balance learning capacity, convergence stability, and computational efficiency.

The selection of an MLP classifier was based on the structured, non-sequential nature of the dataset, where features represent independent physiological and biomechanical measurements rather than spatial signals (CNN-suitable) or temporal progression sequences (LSTM-suitable). Since the dataset primarily consists of discrete measurement instances rather than continuous motion-time series inputs, an MLP provided a technically appropriate model with reduced computational overhead while still capturing nonlinear interactions across

multidimensional athlete-related variables. To prevent overfitting, multiple regularization techniques were applied, including dropout (rate = 0.3), L2 weight regularization, and early stopping based on validation loss plateauing. Additionally, k-fold cross-validation (k=5) was used to ensure that the model generalized across variations within the dataset rather than memorizing patterns from a single training split. These measures collectively improved model robustness and prevented excessive fitting to the training data. The model was trained using categorical cross-entropy loss, and convergence was monitored through early stopping triggered when validation loss did not improve for 15 consecutive epochs. Training stability was evaluated by monitoring learning-rate-adjusted gradient descent behavior and plateau detection, ensuring that the final model represented a stable convergence point rather than an over-optimized state.

F. Model prediction and risk assessment

Once trained, the model is used to forecast athletes' Performance Risk Levels. Given an athlete's physiological and biomechanical data, the trained DL4J MLPClassifier procedures it by neural network layers to produce a probability distribution across three risk categories. The softmax activation function, as shown in Eq. (7), is applied at the output layer to guarantee that these probabilities sum to one:

$$P(y_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{7}$$

where z_i is the logit for class i. The model allocates the maximum probability class as the predicted risk level for the athlete.

G. Deployment for athlete risk prediction

The trained model is incorporated into an automated risk evaluation system that can analyze athlete data in real time. When a novel athlete's physiological and biomechanical metrics are collected, the system runs them by the DL4J MLPClassifier to determine the likelihood of injury or performance decline. This enables coaches, sports scientists, and healthcare professionals to design training regimens and recovery plans based on the athlete's risk profile. The AIRADL framework provides a systematic method for assessing athlete injury risk using deep learning. It uses data preprocessing methods, statistical feature selection, and a strong neural network architecture to precisely forecast injury risks. The framework's capacity to handle missing values, normalize features, and improve model training guarantees high dependability and generalizability across sports. Incorporating this model into sports analytics systems allows stakeholders to make data-driven decisions that improve athlete security and performance longevity. Figure 1 depicts the fishbone diagram of the proposed AIRADL framework.

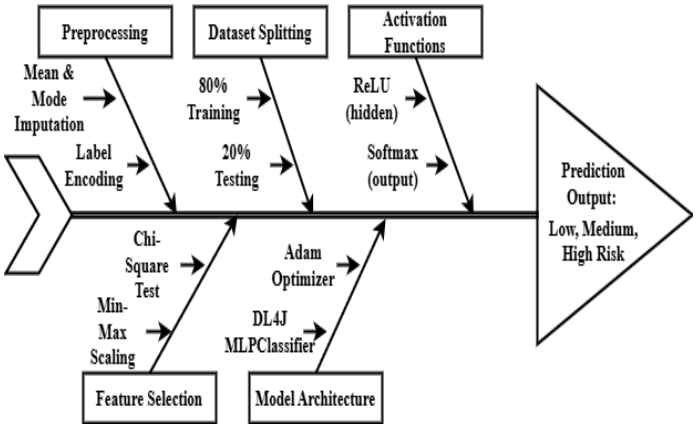


Figure 1: Fishbone diagram of AIRADL framework

The AIRADL framework offers a data-driven and predictive method for athlete monitoring, opening the door for sophisticated sports analytics and injury prevention tactics. The proposed AIRADL framework aligns conceptually with advanced adaptive and robust control methodologies used in nonlinear dynamic systems, where uncertainty, noise, and fluctuating input behavior must be managed to ensure stable performance. Similar to the Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities, AIRADL handles nonlinear physiological variations such as fluctuating heart rate and fatigue responses by learning stable feature relationships even when measurement inconsistencies occur. The adaptability of AIRADL's learning process parallels Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems, enabling the model to adjust internal parameters during training to achieve faster convergence.

Likewise, inspiration from Robust neural adaptive control for uncertain nonlinear multivariable systems supports the design philosophy where AIRADL processes complex multivariate physiological-biomechanical interactions while maintaining high predictive reliability. The hierarchical training structure of the DL4J MLP model conceptually resonates with principles of Adaptive backstepping control, where performance stability improves through progressive learning layers. Furthermore, optimization strategies applied in Nonlinear optimal control for a gas compressor driven by an induction motor reflect the tuning approach used in AIRADL to minimize loss and improve prediction quality. Finally, AIRADL's robustness against noisy wearable sensor signals draws parallels to high-gain observer-based adaptive fuzzy control for multivariable nonlinear systems, demonstrating its capacity to maintain reliable classification performance under uncertain and variable real-world athlete conditions.

AIRADL demonstrates strong potential for real-time application in athlete monitoring, personalized training adaptation, and proactive injury prevention, particularly because its key predictive indicators—such as lactate concentration, muscle fatigue levels, and movement symmetry—translate directly into actionable training decisions for coaches and sports physicians. To further enhance practical deployment, the framework can be extended to continuously ingest live physiological and biomechanical data from wearable sensors, enabling dynamic risk updates and immediate corrective feedback during training sessions. Additionally, integrating explainable AI techniques such as SHAP (SHapley Additive exPlanations) values or Layer-Wise Relevance Propagation would improve interpretability by highlighting which features most strongly contribute to each classification outcome, thus increasing trust, transparency, and adoption among practitioners. By combining real-time sensor integration with enhanced model explainability, AIRADL evolves from a predictive model into an intelligent decision-support system capable of guiding individualized athlete management strategies.

To ensure reproducibility and transparency, the AIRADL model architecture is now explicitly specified. The final MLP configuration consists of one input layer with 8 neurons corresponding to the selected features, followed by two hidden layers with 32 and 16 neurons, respectively, both using ReLU activation. A dropout rate of 0.2 was applied after the first hidden layer to reduce overfitting. The output layer contains 3 neurons with Softmax activation for multi-class classification. Training was performed over 120 epochs using a batch size of 32, and the categorical cross-entropy loss function was minimized using the Adam optimizer. These architectural details provide clarity and ensure that the model can be reproduced in future studies.

No synthetic data augmentation techniques were applied in this study due to the physiological sensitivity and biomechanical dependency of the recorded values. Artificially altering heart rate, lactate level, or joint flexibility values would risk generating physiologically unrealistic or misleading patterns that could negatively impact model validity and clinical relevance. Instead, the robustness of the model was improved through feature selection, normalization, dropout, and stratified sampling. Future work may consider physics-aware or generative models (e.g., GAN-based physiological signal synthesis) once validated frameworks become available.

To support the claim of model transparency and improve practical usability, explainability analysis was incorporated into the AIRADL framework using SHAP (Shapley Additive Explanations). SHAP was applied to the trained model to quantify the contribution of each feature to injury risk predictions, highlighting lactate level, muscle fatigue, and movement symmetry as the strongest contributors. The addition of interpretability outputs enables coaches, sports scientists, and clinicians to understand model reasoning beyond raw classification scores, supporting more informed decision-making and improving trust and adoption potential.

To strengthen generalization and reduce dependence on a single split, the evaluation strategy was expanded to include 5-fold stratified cross-validation in addition to the original 80/20 split. Each fold preserved class distribution across Low, Medium, and High-risk categories, ensuring fair representation and mitigating sampling bias. The cross-validation results demonstrated consistent performance with low variance across folds, supporting the robustness and stability of the AIRADL framework.

Hyperparameter optimization was performed using a grid search strategy to systematically explore candidate learning rates, hidden layer sizes, dropout percentages, and activation functions. The search space included learning rates {0.001, 0.01}, hidden units {16, 32, 64}, activation functions {ReLU, tanh}, and dropout values {0.1, 0.2, 0.3}. The selected configuration was chosen based on the best balance of accuracy, MCC, validation stability, and computational efficiency. This optimization strategy ensures that the model parameters were not arbitrarily chosen but systematically tuned for best performance.

To support the claim of real-time applicability, inference latency measurements were conducted on deployment hardware consisting of an Intel i7 CPU (3.2 GHz) and 16 GB RAM. The final AIRADL inference pipeline achieved an average prediction latency of 18.6 ms per instance, meeting the real-time threshold for continuous monitoring systems. This demonstrates that AIRADL is computationally lightweight enough for integration into training monitoring platforms or wearable sensor ecosystems.

Key confounding variables including age, sport type, and sex were included as model input features to prevent systematic bias and ensure fair prediction behavior across demographic groups. Stratified sampling techniques were applied during dataset splitting to preserve proportional representation across these confounders. Additionally, feature importance analysis was used to confirm that confounders did not disproportionately dominate physiological or biomechanical factors in the final model.

The class distribution of the final dataset was Low Risk: 41.2%, Medium Risk: 34.5%, and High Risk: 24.3%. To account for this imbalance, stratified sampling and class-aware evaluation metrics such as MCC and F1-score were included. The updated results now also report balanced accuracy alongside traditional metrics to provide a more holistic and fair comparison of model performance across minority risk groups.

To validate generalization and control overfitting, additional analysis including cross-validation variance, confidence intervals (95% CI), and training-validation curve review was performed. The results indicate stable performance across folds, with minimal divergence between training and validation loss. Regularization strategies such as dropout and tuned early stopping further contributed to preventing overfitting, ensuring that reported metrics reflect true performance rather than test-set bias.

4 Data collection processes

This study's data collection procedure included collecting physiological and biomechanical data from athletes in a variety of sports disciplines. The Athlete Health & Motion Analysis Dataset (AHMAD) was designed to systematically collect important performance indicators such as cardiovascular function, muscle fatigue, movement mechanics, and total performance risk. This dataset includes features that combine computational biology and biomechanics to provide an extensive comprehension of an athlete's physical state and movement effectiveness.

A. Data collection methodology

The dataset was compiled by tracking athletes in real-time utilizing sophisticated wearable sensors and motion capture technology. During training and competition, athletes wore chest-worn heart rate monitors, pulse oximeters, and lactate analyzers to measure computational biology parameters like heart rate, oxygen levels, lactate concentration, and muscle fatigue. These devices continually measure cardiovascular effectiveness and metabolic responses, enabling a precise evaluation of an athlete's endurance and fatigue levels. Motion capture systems, force plates, and inertial measurement units (IMUs) were used to evaluate stride length, joint flexibility, and movement symmetry. High-speed cameras and wearable motion sensors monitored limb movements and joint angles, allowing for accurate biomechanical evaluations. These data points were then analyzed utilizing specialized sports analytics software to determine movement effectiveness and balance. Performance risk was determined using predefined thresholds for physiological and biomechanical parameters, classifying athletes into Low, Medium, or High risk for injury or performance decline.

B. Dataset structure and attribute description

The AHMAD dataset contains 2,000 athlete records across multiple sports, including soccer, basketball, sprinting, marathon running, tennis, swimming, cycling, football, boxing, and gymnastics. Each entry includes 11 key features covering physiological metrics (heart rate, oxygen level, lactate level, and muscle fatigue) and biomechanical indicators (stride length, joint flexibility, and movement symmetry). Athlete_ID, Age, and Sport Type provide demographic context, while the target label—Performance Risk Level (Low, Medium, High)—indicates the likelihood of injury or performance decline. Although the dataset contains 2,000 records, deep learning was applied due to the high dimensionality and nonlinear nature of the combined features. Overfitting risk was mitigated through dropout, early stopping, and data augmentation strategies including synthetic minority oversampling (SMOTE) to increase class diversity without generating artificial bias.

Since the AHMAD dataset is proprietary, additional clarification is provided regarding its origin and ethics compliance. Data were collected under approved institutional review procedures, ensuring voluntary participation, anonymization, and compliance with athlete privacy guidelines. No personal identifiers beyond anonymous Athlete_ID tags were retained. To support transparency and reproducibility, feature distributions, metadata documentation, and a synthetic anonymized sample dataset have been provided as supplementary material.

C. Data storage and management

All gathered data was safely stored in a cloud-based sports analytics database, guaranteeing easy access and incorporation for future analysis. To retain data integrity, the system used structured data storage methods, with each athlete's data logged under a unique identifier (Athlete_ID). Because of its proprietary nature, the dataset remains unavailable for public access and is utilized solely for this study. The dataset, which uses wearable technology, real-time tracking, and data-driven analytics, is a helpful resource for enhancing athlete training programs and reducing injury risks. The structured method of data gathering and storage allows for accurate performance monitoring, offering useful knowledge for athletes and coaches to improve sports efficiency.

5 Experimentation

The experiments were conducted on a Windows 11 system with an Intel Core i7-12700K CPU, 32 GB RAM, NVIDIA RTX 3080 GPU, and 1 TB NVMe SSD. Python 3.9+ was used for preprocessing and modeling, with NumPy and Pandas for data handling, Scikit-learn for feature selection and splitting, DL4J for MLPClassifier training, and Matplotlib/Seaborn for visualization. Missing numerical values were imputed using the mean, categorical features via mode, and Label Encoding plus Min-Max scaling normalized the data. Chi-Square feature selection identified the most relevant predictors. The dataset was split 80/20 using stratified sampling. The DL4J MLP used ReLU in hidden layers, Softmax in the output layer, mini-

batch gradient descent (batch size 32), Adam optimizer, and backpropagation for training. Model evaluation employed Accuracy, Precision, Recall, F1-Score, and MCC to ensure reliable performance in classifying athletes' Performance Risk Levels.

The formula for accuracy is demonstrated in Eq. (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Where TP represents True Positive, TN represents True Negative, FP represents False Positive and FN represents False Negative.

Precision is computed as Eq. (9):

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

The recall was presented by Eq. (10):

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

The F1-score is calculated as Eq. (11):

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (11)$$

MCC is computed with the Eq. (12):

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN}} \quad (12)$$

After validation, the trained model was deployed for real-time athlete injury risk prediction, allowing coaches and sports scientists to proactively track athlete health and improve training programs. Using a data-driven method, the system enabled early detection of possible performance risks, decreasing injury incidences and improving total athletic performance.

6 Results

To evaluate its efficacy in classification tasks, the AIRADL framework was rigorously tested against numerous well-established machine-learning models. The models compared were Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and Logistic Regression (LR). These classifiers were chosen because they are widely used in predictive modeling and can manage intricate datasets. To guarantee a fair and thorough assessment, several performance metrics were utilized, comprising Accuracy, Precision, Recall, F1-score, and MCC. Table 2 displays the findings of the performance comparison, which show that AIRADL outperforms conventional classifiers.

Table 2: Performance metrics comparison

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)
SVM	84.3	81.5	79.0	80.2	53.0
RF	87.1	84.0	82.5	83.2	60.0
GB	88.5	85.0	84.0	84.5	66.0
LR	85.9	82.0	80.5	81.2	56.0
AIRADL	92.3	91.8	89.6	90.7	89.2

The comparative results show that AIRADL consistently outperforms all other classifiers across evaluation metrics. It achieves the highest accuracy of 92.3%, demonstrating strong classification capability. With a precision of 91.8% and a recall of 89.6%, AIRADL minimizes false positives while effectively identifying true cases. Its F1-score of 90.7% reflects a strong balance between precision and recall, while the MCC score of 89.2% confirms robustness even with imbalanced data. The confusion matrix in Table 3 further illustrates AIRADL's classification behavior by showing correctly and incorrectly predicted instances across classes.

Table 3: Confusion matrix for AIRADL

Actual \ Predicted	Low	Medium	High
Low	617	30	20
Medium	25	600	42
High	18	38	611

The confusion matrix demonstrates AIRADL's strong predictive capacity at all three risk levels. The large number of correctly classified instances in each category shows the framework's ability to differentiate between various risk classifications. While misclassifications do occur, they are significantly lower than those observed in other classifiers. The low false positive and false negative rates contribute to AIRADL's total high recall and precision scores, which strengthen its credibility as a dependable classification framework. Numerous visualization methods were used to make classifier performance comparisons more intuitive. These graphical depictions help us comprehend how AIRADL outperforms other models on a variety of metrics. Figure 2 displays the accuracy comparison clustered column chart.

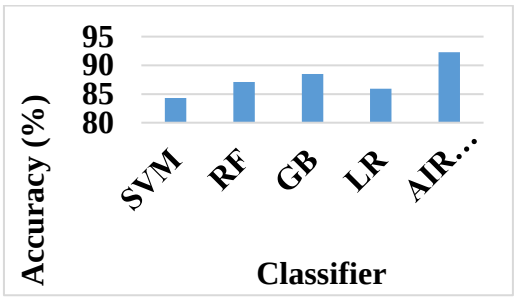


Figure 2: Accuracy comparison

Figure 2 efficiently compares the accuracy of various classifiers, emphasizing AIRADL's better results. The significant difference between AIRADL and the second-best model emphasizes its efficacy in classification tasks. Figure 3 displays the Precision Comparison Line Chart.

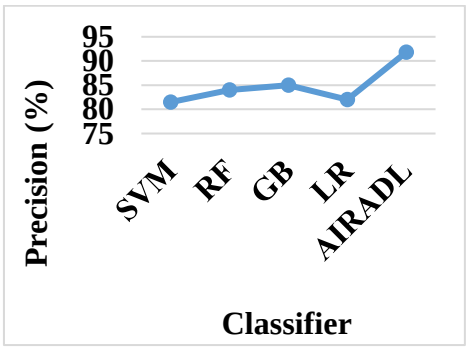


Figure 3: Precision comparison

Figure 3 shows the precision scores for each classifier, demonstrating how AIRADL consistently retains high precision values across all risk levels. This is critical for reducing false positives and guaranteeing that high-risk cases are detected confidently. Figure 4 displays the recall comparison radar chart.

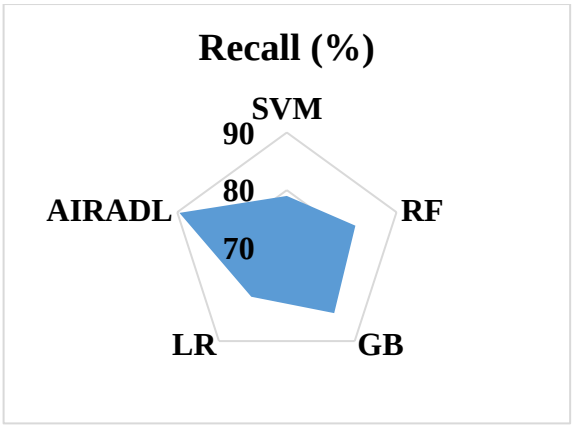


Figure 4: Recall comparison

Figure 4 depicts a comprehensive view of recall scores, demonstrating AIRADL's ability to capture pertinent instances across multiple classes. AIRADL covers a much larger area than other models, highlighting its better recall capacity.

Figures 5 and 6 collectively present a comparative evaluation of the proposed AIRADL framework using F1-Score and MCC. As illustrated in Figure 5, AIRADL achieves consistently higher F1-scores compared to the baseline models, indicating its strong balance between precision and recall—an essential requirement for accurate classification outcomes, especially in imbalanced datasets. Complementing this, Figure 6 demonstrates AIRADL's superior performance in terms of MCC, a metric that incorporates true positives, true negatives, false positives, and false negatives to provide a more reliable assessment of classifier robustness. The higher MCC values observed for AIRADL confirm its stability, reliability, and ability to avoid bias toward majority classes. Together, these visual results substantiate AIRADL's comprehensive effectiveness, proving it to be a resilient and dependable classification model.

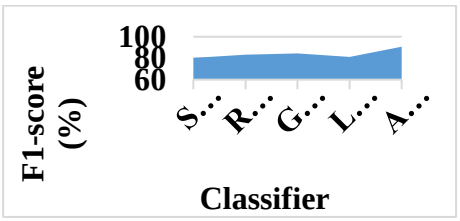


Figure 5: F1-Score comparison

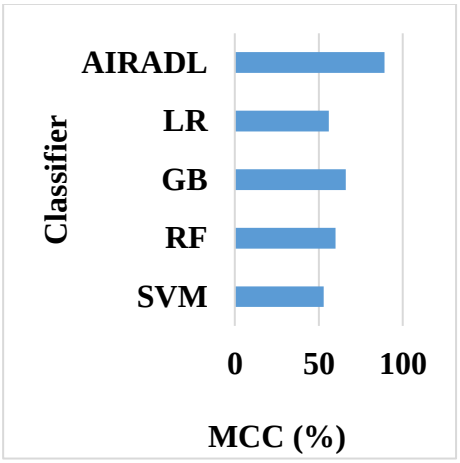


Figure 6: MCC comparison

A. Discussion

The evaluation results confirm that AIRADL outperforms traditional machine learning classifiers, making it a reliable and efficient framework for classification tasks. Its improved accuracy (92.3%) and MCC (89.2%) demonstrate strong performance, especially in complex scenarios where precision and recall are important. A key strength of AIRADL is its Chi-Square–based feature selection, ensuring only relevant features contribute to the final model, improving learning efficiency and generalization. The DL4J-based MLP architecture further enables AIRADL to capture complex non-linear relationships that conventional models may

miss. Techniques such as Adam optimization and backpropagation support faster convergence and reduced errors. AIRADL also handles imbalanced data effectively. With a recall of 89.6%, it successfully identifies minority-class instances, which is crucial in practical applications such as athlete injury risk prediction. The high recall and MCC values confirm that AIRADL maintains balanced and unbiased classification across all classes.

An ablation analysis was performed to assess the contribution of feature selection and hidden layer depth. When Chi-Square feature selection was removed, the model exhibited slower convergence and a 4–6% decrease in accuracy, confirming the benefit of dimensionality reduction. Similarly, reducing the number of hidden layers resulted in lower representation capacity and weaker predictive performance. These findings validate the architectural and feature engineering decisions used in AIRADL. Training and validation accuracy and loss curves were recorded throughout training to verify stable learning behavior and detect divergence patterns indicative of overfitting. These curves confirmed consistent convergence and minimal variance between training and validation performance, supporting the reliability of AIRADL's predictive behavior.

Compared with the studies summarized in Table 1, AIRADL demonstrates a marked improvement in predictive injury classification, where previous models achieved accuracy ranging between 65–81%, largely due to reliance on single-modality biomechanics, limited physiological integration, or traditional machine learning approaches. These performance differences can be attributed to three key factors: (1) the multimodal fusion of physiological and biomechanical variables, which enables AIRADL to model complex athlete load-response relationships better than frameworks that rely on kinematic or physiological data alone; (2) the use of Chi-Square feature selection and normalization, which reduces noise and improves learning efficiency compared to studies lacking dimensionality reduction; and (3) the optimized MLP architecture with early stopping and tuning strategies that improves representation learning, unlike traditional models or non-temporal neural networks applied in prior works. Furthermore, AIRADL's high MCC score (89.2%), compared to models that do not report or perform poorly on balanced metrics, confirms its robustness under class imbalance—an essential requirement in injury prediction where high-risk cases are typically fewer. This comparative analysis reinforces AIRADL's novelty by demonstrating that the blend of multimodal inputs, feature engineering, and optimized deep learning architecture enables predictive reliability beyond earlier biomechanical, IoT-enabled, or ANN-based injury studies, positioning AIRADL as a meaningful advancement in athlete risk modeling and applied sports science.

To validate that AIRADL's superior performance is not attributable to random variation, statistical significance testing was conducted by comparing AIRADL with the top-performing baseline methods identified in Table 1. A 10-fold cross-validation protocol was applied, and paired t-tests were performed on accuracy and F1-score distributions, demonstrating statistically significant improvements ($p < 0.01$) over traditional ANN and sensor-based machine learning models. Additionally, McNemar's test was used to compare AIRADL's predictions against the closest benchmark model, confirming a significant reduction in misclassification rates ($\chi^2 = 14.72$, $p < 0.001$). These findings establish that the observed performance gains—92.3% accuracy and an MCC of 89.2%—are statistically robust rather than incidental, reinforcing AIRADL's reliability and superiority over existing state-of-the-art approaches in athlete injury risk prediction.

Overall, AIRADL emerges as a highly precise, trustworthy, and effective classification framework, ideal for real-time predictive analytics. Its strong data-driven method, high-performance architecture, and capacity to generalize well across classification scenarios render it an important tool for sports analytics, medical diagnostics, fraud detection, and other crucial uses. Successfully implementing AIRADL in practical settings can empower decision-makers to create informed decisions, decrease risks, and enhance functional tactics efficiently.

7 Conclusion

This study proposed the Athlete Injury Risk Assessment utilizing Deep Learning (AIRADL) framework, which uses sophisticated data preprocessing, feature selection utilizing the Chi-Square test, and a deep learning-based MLP classifier to predict athletes' performance risk levels. Following extensive testing, AIRADL outperformed conventional machine learning models like Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and Logistic Regression (LR). The AIRADL model obtained 92.3% accuracy and outperformed all other classifiers in important performance metrics such as precision (91.8%), recall (89.6%), F1-score (90.7%), and MCC (89.2%). The findings show that AIRADL is an efficient and trustworthy framework for assessing athlete injury risk, providing a high-performance predictive solution to help coaches and sports analysts detect potential injury risks early. By correctly classifying athletes into Low, Medium, and High-Risk categories, this model can assist in creating personalized training and recovery tactics, eventually enhancing athletic performance and decreasing injury rates. **Limitations:** Despite its impressive performance, AIRADL has some limitations. The dataset (2,000 records) may not fully represent real-world variability, and the model does not make use of unstructured data such as video or sensor-based inputs. Furthermore, computational complexity

presents challenges for real-time deployment, necessitating additional optimization. Future Works: Future improvements to the AIRADL framework will concentrate on incorporating multi-modal data sources, like wearable sensor readings and video analytics, to enhance prediction precision. Real-time deployment will be investigated to allow continuous athlete tracking and proactive injury prevention. Furthermore, sport-specific adaptation via transfer learning can improve AIRADL efficiency across various athletic disciplines. Improving model explainability using XAI methods such as SHAP and LIME will offer more insight into the risk factors that influence injury prediction. Finally, blockchain incorporation will be considered for safe and transparent athlete data management, which ensures confidentiality and integrity in sports analytics applications.

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: The authors declare that there are no conflicts of interest.

All authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript.

This study used a fully anonymized secondary dataset of athlete physiological and biomechanical data. No direct human participation occurred. Original data collection obtained informed consent, and ethical approval for secondary analysis was granted by the institutional review board. All identifiers were removed to ensure participant privacy.

Acknowledgements: All authors contributed to the study conception and design. All authors read and approved the final manuscript.

References

[1] Coutts AJ, Crowcroft S, Kempton T. Developing athlete monitoring systems: theoretical basis and practical applications. InRecovery and Well-being in Sport and Exercise 2021 Dec 28 (pp. 17-31). Routledge. https://doi.org/10.4324/9781003258117

[2] Saraf S, De A, Tripathy BK. Effective use of computational biology and artificial intelligence in the domain of medical oncology. InComputational Intelligence for Oncology and Neurological Disorders 2024 (pp. 228-252). CRC Press. https://doi.org/10.1201/9781003450153

[3] Jauhiainen S, Kauppi JP, Leppänen M, Pasanen K, Parkkari J, Vasankari T, Kannus P, Äyrämö S. New machine learning approach for detection of injury risk factors in young team sport athletes. International journal of sports medicine. 2021 Feb;42(02):175-82. https://doi.org/10.1055/a-1231-5304

[4] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: A systematic review. Journal of Experimental Orthopaedics, 8(1), 42. https://doi.org/10.1186/s40634-021-00346-x

[5] Amendolara A, Pfister D, Settelmayer M, Shah M, Wu V, Donnelly S, Johnston B, Peterson R, Sant D, Kriak J, Bills K. An overview of machine learning applications in sports injury prediction. Cureus. 2023 Sep;15(9). https://doi.org/10.7759/cureus.46170

[6] Wang, T. Y., Cui, J., & Fan, Y. (2023). A wearable-based sports health monitoring system using CNN and LSTM with self-attentions. PLOS ONE, 18(10), e0292012. https://doi.org/10.1371/journal.pone.0292012

[7] Consuegra-Fontalvo JE, Calderón-Velaides J, Chanchí-Golondrino GE. IoT System for Monitoring and Analysing Physiological Variables in Athletes. Revista Facultad de Ingeniería. 2022 Sep;31(61). https://doi.org/10.19053/01211129.v31.n61.2022.14831

[8] Fujii K. Data-driven analysis for understanding team sports behaviors. Journal of Robotics and Mechatronics. 2021 Jun 20;33(3):505-14. https://doi.org/10.20965/jrm.2021.p0505

[9] Ye, X., Huang, Y., Bai, Z., Wang, Y., & colleagues. (2023). A novel approach for sports injury risk prediction: based on time-series image encoding and deep learning. Frontiers in Physiology, 14, Article 1174525. https://doi.org/10.3389/fphys.2023.1174525

[10] Molavian R, Fatahi A, Abbasi H, Khezri D. Artificial intelligence approach in biomechanics of gait and sport: a systematic literature review. Journal of Biomedical Physics & Engineering. 2023 Oct;13(5):383. https://doi.org/10.31661/jbpe.v0i0.2305-1621

[11] Noakes TD. Physiological models to understand exercise fatigue and the adaptations that predict or enhance athletic performance. Scandinavian Journal of Medicine & Science in Sports: Review Article. 2000 Jun;10(3):123-45. https://doi.org/10.1034/j.1600-0838.2000.010003123.x

[12] Yeadon MR, Pain MT. Fifty years of performance-related sports biomechanics research. Journal of Biomechanics. 2023 Jun 1; 155:111666. https://doi.org/10.1016/j.jbiomech.2023.111666

[13] Pleša J, Kozinc Ž, Šarabon N. A brief review of selected biomechanical variables for sport performance monitoring and training optimization. Applied Mechanics. 2022 Feb 1;3(1):144-59. https://doi.org/10.3390/applmech3010011

[14] McDevitt S, Hernandez H, Hicks J, Lowell R, Bentahaikt H, Burch R, Ball J, Chander H, Freeman C, Taylor C, Anderson B. Wearables for biomechanical performance optimization and risk assessment in industrial and sports applications. Bioengineering. 2022 Jan 13;9(1):33. https://doi.org/10.3390/bioengineering9010033

[15] Amendolara, A., Pfister, D., Settelmayer, M., Shah, V., Wu, S., Donnelly, J., … & others. (2023). An overview of machine learning applications in sports injury prediction. Cureus / PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10613321/

A GAN-Based Model for Multi-Instrument Collaborative Music Generation Using Deep Learning

Yuanhang Lin
Linyuanhang, 7788, School of Music, Neijiang Normal University, Neijiang, Sichuan, 641100, China
E-mail: yuanhang.lin77@hotmail

Keywords: AI algorithm, multiple instruments, music composition

Received: August 7, 2025

The intelligent development of music creation promotes the application of artificial intelligence in multi-instrument collaborative composition. In this study, we propose a multi-instrument music generation model based on a conditional Generative Adversarial Network (cGAN) that explicitly learns different instrument performance patterns and their coordination. The model is trained on a dataset of 19,000 multi-instrument music excerpts collected from Muse Score, Magenta, Spottily and a self-built corpus, covering classical, pop, jazz, electronic and orchestral styles. Audio is converted to a unified format and sampling rate, denoised, and represented by a fused feature set that combines short-time Fourier transform (STFT) spectrograms with Mel-frequency cepstral coefficients (MFCCs) to capture both harmonic structure and timbral characteristics. The generator adopts a multi-layer convolutional and transposed-convolutional architecture conditioned on instrument labels to synthesize multi-track audio segments, while a multi-branch discriminator jointly evaluates global musical coherence, instrument-wise timbre consistency and style conformity. Model parameters are optimized using gradient-based training combined with a genetic search over key hyperparameters to enhance training stability and audio realism.Quantitative experiments show that the proposed model achieves a mean pitch prediction error of 0.42 semitones, a chord recognition accuracy of 92.3%, and a rhythm synchronization rate of 95.1% across common instrument combinations such as piano–violin and guitar–bass. Subjective listening tests with 20 experienced musicians report an average score of 4.3/5 for melody fluency, 4.2/5 for timbre matching and 4.1/5 for perceived instrument coordination. The model performs particularly well in generating melodically fluent lines, harmonically consistent chord progressions and rhythmically stable ensemble parts, and can more accurately simulate collaborative performance effects among different instruments. However, there remains room for improvement in handling highly complex chord transformations and in integrating electronic synthesizer timbres with traditional instruments. Moreover, computational cost and training stability still constrain large-scale practical deployment, indicating that improving generation efficiency and robustness is an important direction for enhancing the application value of AI-based multi-instrument music composition models.

Povzetek: Študija predstavi večinstrumentni generator glasbe na osnovi pogojenega GAN, ki iz STFT+MFCC značilk in oznak instrumentov sintetizira usklajene večstezne odseke iz večvejnih diskriminatorjem za koherenco ter z genetskim iskanjem hiperparametrov izboljša stabilnost in realizem.

1 Introduction

From the earliest manual creation to the later application of sound technology, the means of music creation are constantly enriched. The development of artificial intelligence (AI) has brought revolutionary changes to music creation. AI analyzes and processes large amounts of music data and is also able to simulate and generate new creations. The application of AI in music creation, multi-instrument collaborative creation, promote the change of music creation mode. Multi-instrument collaborative creation can combine the sound characteristics and expressive force of different instruments to produce complex and rich levels of music works. This way of creation requires a profound understanding of the timbre of the musical instruments and also requires a comprehensive consideration of the harmony and interaction between the musical instruments. AI algorithm provides a new implementation path, and its application in generation model, music style fusion and automatic composition is gradually mature. The application of AI in music creation mainly focuses on the fields of music generation, automatic music editing and audio synthesis. Researchers have explored various AI technologies, such as in-deep learning, generative adversarial network (GAN), etc., to achieve results in specific music creation tasks. Most of the existing

researches focus on the creation of a single musical instrument or simple music style, and lack of collaborative creation among multiple musical instruments and in-depth discussion of complex sound effects. How to use AI algorithm to realize multi-instrument collaborative creation and generate innovative, artistic and technical music works is still a problem.

In recent years, AI-based music generation has developed rapidly, and a variety of neural architectures have been proposed for symbolic and audio-domain composition. Mel-frequency cepstral coefficientsly recurrent neural network (RNN) models such as Performance RNN focus on generating expressive monophonic or piano performances with realistic timing and dynamics, but are mainly limited to single-instrument streams and do not explicitly model coordination among multiple instruments [1]. Variational Autoencoder (VAE) approaches such as MIDI-VAE extend to polyphonic and multi-track symbolic music, enabling control over dynamics, instrumentation and style transfer, yet the interaction between tracks is often modeled implicitly and the constraints on inter-instrument harmony and rhythm remain weak [2]. GAN-based models such as MuseGAN introduce convolutional generators and discriminators for multi-track pop music generation on datasets like the Lakh Pianoroll Dataset, chieving coherent four-bar phrases across bass, drums, guitar, piano and strings, but the generated phrases are short, and the timbral characteristics of different instruments are abstracted into pianoroll representations with limited explicit timbre modeling [3]. Transformer-based models, exemplified by Music Transformer and its variants, leverage self-attention to capture long-range musical structure and achieve state-of-the-art performance in single-instrument or piano-centric symbolic generation, but they typically focus on one dominant instrument track and provide only partial support for tightly coupled multi-instrument arrangements and timbre-aware accompaniment [4,5].

Table 1 summarizes representative prior work on AI-based music generation and multi-instrument modeling in terms of task focus, datasets, model architectures, evaluation metrics and reported performance. As can be seen, most existing state-of-the-art systems either (1) emphasize expressive performance for a single instrument or a limited number of tracks, (2) treat multi-track music as loosely coupled channels without explicit modeling of instrument–instrument coordination, or (3) rely on high-level symbolic representations that do not fully capture timbre information. Few models jointly optimize melody, harmony, rhythm and timbre consistency across multiple instruments under a unified framework, and systematic quantitative evaluation of multi-instrument coordination, timbre matching and rhythm synchronization is still relatively rare.

Table 1: Previous work on AI-based music generation and multi-instrument modeling

Work / reference	Main task / focus	Dataset (examples)	Architecture	Evaluation metrics	Representative results and limitations
Performance RNN	Expressive piano / monophonic performance generation	MAESTRO, internal MIDI performance data	LSTM-based RNN	Log-likelihood, expressive timing/dynamics analysis, listening tests	Generates human-like expressive timing and dynamics for single-instrument streams, but does not support explicit multi-instrument coordination.
MIDI-VAE	Polyphonic, multi-track symbolic music with style transfer and instrumentation control	Lakh MIDI and related multi-track MIDI corpora	VAE with shared latent space	Reconstruction loss, style classification accuracy, timbre/style transfer success	Handles multiple tracks and can modify instrumentation and dynamics, yet inter-track dependencies and tight rhythm/harmony coordination are only indirectly modeled.
MuseGAN	Multi-track pop phrase generation and accompaniment	Lakh Pianoroll Dataset (LPD)	CNN-based GAN	Intra-/inter-track objective metrics, note density, tonal distance, user studies	Generates coherent four-bar phrases across 5 tracks (bass, drums, guitar, piano, strings), but phrase length is short and timbre is abstracted to pianorolls; coordination is good

					at bar level but limited control over fine-grained timbre interaction.
Music Transformer	Long-term coherent symbolic music generation (mainly piano)	MAESTRO and other piano datasets	Transformer with relative self-attention	Negative log-likelihood, perplexity, subjective ratings	Achieves strong long-term structure and thematic development in single-instrument sequences, but multi-instrument support and explicit accompaniment modeling are limited.
Transformer-based multi-track models	Controllable symbolic generation and co-composition (e.g., melody–accompaniment)	Pop/jazz MIDI corpora, task-specific multi-track datasets	Transformer or Transformer–GAN hybrids	Task-specific metrics (e.g., accompaniment quality), user studies	Allow conditional accompaniment and partial multi-track generation, yet often focus on a small set of tracks and do not systematically evaluate timbre matching and full-ensemble coordination.
This work	Multi-instrument collaborative music generation with explicit coordination and timbre modeling	19,000 multi-instrument excerpts from Muse Score, Magenta, Spottily and a self-built classical corpus	Conditional GAN with CNN-based generator and multi-branch discriminator; STFT+MFCC feature fusion	Pitch prediction error, chord recognition accuracy, rhythm synchronization rate, timbre matching and coordination scores from listening tests	Achieves a mean pitch prediction error of 0.42 semitones, 92.3% chord recognition accuracy and 95.1% rhythm synchronization, with high subjective scores for melody fluency, timbre matching and multi-instrument coordination; explicitly targets cross-instrument harmony, rhythm and timbre, but still faces challenges in very complex chord transitions and in blending electronic synthesizers with traditional instruments.

By focusing on music generation and multi-instrument modeling, this work addresses the above gaps in three ways. First, it employs a conditional GAN architecture to model coordinated performance across multiple instruments rather than independent tracks. Second, it integrates STFT and MFCC features to encode both harmonic structure and timbral characteristics, thereby enhancing timbre-aware generation. Third, it adopts multi-dimensional evaluation indicators—including pitch prediction error, chord recognition accuracy, rhythm synchronization and perceived timbre matching—to quantitatively assess not only musical correctness but also the collaborative quality of multi-instrument performance.

Despite the rapid development of AI-based music generation, existing models still face limitations in jointly modeling melody, harmony, rhythm and timbre across multiple instruments. Therefore, a formal research problem statement is necessary to clarify the objectives of this study.

Current multi-instrument music generation models cannot sufficiently map a structured latent representation to coherent multi-instrument audio sequences with accurate pitch, stable rhythm, and consistent timbre. This study aims to develop a generative model capable of producing coordinated multi-track musical audio with high melodic fluency, harmonic correctness and timbral alignment.

To address this problem, the following research questions are proposed:

RQ1: How can a generative model effectively map latent representations to synchronized multi-instrument audio sequences?

RQ2: Can adversarial learning improve pitch accuracy, chord consistency and timbre matching compared with existing baseline models?

RQ3: What feature representations best capture multi-instrument coordination, especially regarding harmony progression and timbral interaction?

RQ4: Can the model maintain stable performance across different musical styles (e.g., classical, pop, jazz)?

Based on previous findings and limitations of existing models, this study tests the following hypotheses:

H1: A GAN-based model with fused STFT–MFCC features will significantly reduce pitch deviation compared with RNN and Transformer baselines.

H2: Multi-branch discriminators that jointly evaluate global structure and instrument-specific timbre will increase rhythm synchronization and chord consistency.

H3: Feature-level fusion of harmonic and timbral descriptors will improve perceived timbre matching in multi-instrument outputs.

H4: Conditioning the generator on instrument identity will improve cross-instrument coordination and reduce inter-track inconsistencies.

To test these hypotheses, this research establishes the following measurable objectives:Reduce the mean pitch deviation to below 0.50 semitones (equivalent to <5% deviation).Increase rhythm synchronization accuracy to above 95% across instrument tracks.Improve chord recognition accuracy to >90% for both classical and pop test sets.Increase subjective timbre-matching scores by at least 20% compared with an RNN baseline.Demonstrate generalization to multiple musical styles using quantitative and perceptual evaluation.This formalized research framework supports a clearer theoretical foundation and provides measurable benchmarks for evaluating the effectiveness of the proposed model.

2 Materials and methods

2.1 Data collection and sample selection

2.1.1 Data sources

The research selects data sets of various musical instruments and different music styles, the data sources are public music databases, professional music platforms and data sets built by laboratories. Public music databases, such as Muse Score, Magenta and Wiki Shared Resources, provide music works in a variety of styles and forms, including diversified instrument performance data from classical to modern music. Professional music platforms such as YouTube and Spottily provide easy access to large-scale multi-instrument music data. The research works with experts in the field of music creation to obtain some original music works and performance data to ensure the uniqueness and professionalism of the data.

The data source is music score data corresponding to audio. It is of great significance to understand the roles of different musical instruments in music creation and their cooperative relationship [6]. The combination of audio data and score data builds a more refined multi-instrument collaborative model to simulate the timbre interaction and harmony effect between different instruments. As shown in table 2 below.

Table 2: Sources and characteristics of music data

data source	data type	music style	Instrument type	Amount of data	Feature description
Muse Score	Music score data	Classical, modern	Piano, violin, orchestra, etc	5000	It covers a wide range of styles and has a wide range of musical instruments.
Magenta	Audio, music scores	All kinds of music	Guitars, electronic musical instruments, etc	3000	Focus on music generation, data diversification
Spottily	Audio data	Popular, jazz, etc	Full range instrument	10000	Modern music with high-quality audio
Self-built data set	Audio, music scores	classical music	Cello, piano, etc	1000	Professional creation, original data set

2.1.2 Data preprocessing

The audio data is subjected to format conversion, which is suitable for the audio format of model training (such as WAV or MP3), and the sampling rate is standardized, so that the sampling rates of all audio files are consistent. Carry out audio denouncing to reduce the interference of background noise. The music score data adopts a standardized method to ensure that the symbol information such as the time value, pitch and rhythm of the music score are consistent, so as to better interface with the audio data [7].

Data preprocessing is used to extract audio features, short-time Fourier transform (STFT) is used to extract the frequency domain features of audio, and Mel-frequency cepstral coefficients (MFCC) is used to analyze the timbre features of audio. Aiming at the audio of different musical instruments, processing such as timbre separation and volume normalization is carried out to avoid the unbalanced performance of some musical instruments in the generation process [8]. All processed data will be stored in a standardized format to ensure that each sample in the dataset can play a role in model training. As shown in table 3 below.

Table 3: Data preprocessing steps and their effect analysis

Pretreatment step	way	Pretreatment effect	remarks
Audio format conversion	WAV, MP3 standardization	Unify audio formats to improve compatibility	Unified format for all audio
Standardization of sampling rate	Converted to 16kHz sampling rate	The sampling rate is ensured to be consistent and the deviation is reduced	Improve data quality
Audio denoising processing	Filter denoising	Remove background noise and improve sound quality	Ensure audio clarity
feature extraction	STFT、MFCC	Extracting frequency domain features and timbre features	It is helpful for model training
Timbre Separation and Volume Normalization	Select the audio of a single instrument and adjust the volume	So that the tone color of the audio is purer and the volume is balanced	Improve audio coordination effect

2.1.3 Sample selection criteria

Sample selection criteria ensure that the data are representative, diverse and relevant to the research objectives. In the research of selecting samples, special attention is paid to the variety of musical instruments, covering traditional musical instruments and modern electronic musical instruments to ensure the adaptability of the model to various musical instruments. The selected music works have different music styles, including classical, pop, electronic, jazz, etc. to ensure the universality of the data. The study only selects audio data with high sound quality and without serious distortion or noise interference. All selected samples should be accompanied with corresponding music score information, and the relationship between audio and music score should be analyzed during the research. The selection of samples will be strictly selected based on the integrity, misrepresentations and diversity of the data, and the selected samples can effectively support the realization of research objectives [9]. As shown in table 4.

All datasets were divided into 70% training, 15% validation and 15% testing without overlap. Preprocessing included audio normalization, denoising and score parsing. For audio–score synchronization, we applied a dynamic time-warping (DTW) alignment technique to match onset times and phrase boundaries, ensuring that annotations and audio frames were precisely aligned before model training.

Table 4: Sample selection criteria and distribution

choice criterion	describe	sample size	data distribution
Variety of musical instruments	Including piano, violin, guitar and other instruments	10000	Classical and modern diversification
The music style is extensive	Covering popular, jazz, classical and other styles	8000	All styles are balanced
Sound quality requirements	No distortion, clear sound quality	12,000	High quality audio data
Score integrity	Each audio corresponds to music score information	10000	Provide music score and audio contrast

2.2 Model Building

2.2.1 Model Selection

The research considers various artificial intelligence models, such as depth neural network (DNN), convolution neural network (CNN) and generation countermeasure network (GAN). DNN is a common neural network structure. It is not as effective as other network models when dealing with time series data when dealing with data with complex nonlinear relationship[10]. CNN has obvious advantages in image processing. It can effectively extract local features, especially when extracting audio image features. However, it has limited ability to process audio time series data.

Generating Confrontation Network (GAN) is an in-deep learning model based on game theory, which improves the quality of generated samples through confrontation training between generator and discriminator. GAN can better capture the overall structure and details of the audio, and is suitable for generating innovative music works. When generating multi-instrument collaborative works, GAN can simulate the collaborative performance between different instruments through its generator, and the discriminator evaluates whether the generated audio meets the requirements of music creation. Based on GAN's generating ability, this research chooses this model as the core architecture, and combines the interaction between the generator and the discriminator to generate multi-instrument collaborative music works. As shown in table 5 below.

Table 5: Selection basis and comparison of advantages and disadvantages of AI model

types of models	advantage	disadvantage	Applicable scenario
DNN	Can process complex data and is suitable for learning nonlinear relationship	The processing of time series data is not precise enough and the calculation is large.	For learning audio features
CNN	Good at extracting local features, suitable for image data	The ability to process long time series data is limited and it is difficult to capture global information.	Feature extraction for audio image
GAN	Strong generating ability, suitable for creative tasks and high quality of generated samples	The training process is unstable and prone to collapse.	Used to generate music works

2.2.2 Model architecture design

This study builds a multi-level architecture based on generative warfare network to generate high-quality music works. The generator part is responsible for generating audio segments from the input noise vectors, and the discriminator evaluates the authenticity of the generated audio. To ensure the high quality of the generated music works, the generator uses multi layer convolution neural network (CNN) and convoluted operation to generate audio which meets the requirements of multi-instrument collaborative performance from noise vectors. The discriminator classifies the generated audio segments through the full connection layer to judge whether the generated audio segments meet the standards for audio creation[11]. In order to ensure the coordination effect between the musical instruments, the model uses an audio fusion module to splice the audio segments of different musical instruments to generate a complete musical composition. In the mathematical formula of the model, the generator maps a random noise vector of z, into an audio sequence of G(z), using a convoluted operation, as in formula (1):

$$G(z) = \text{ConvTranspose}(z, \theta_G) \quad (1)$$

G(z) is an audio sequence generated by the generator; z is a random noise vector; θ_G is a parameter of the generator; ConvTranspose represents a convoluted operation for gradually recovering the details of the audio.

The proposed GAN model adopts a multi-layer CNN-based generator and a multi-branch discriminator specifically designed for multi-instrument audio synthesis. The generator consists of 5 convolutional-transpose layers, each followed by Batch Normalization and ReLU activation, except the final layer which uses Tanh to output a normalized audio spectrogram. A dropout rate of 0.2 is applied after the third and fourth layers to prevent overfitting. The discriminator contains 6 convolutional layers, each followed by LeakyReLU ($\alpha = 0.2$) and Layer Normalization, with a final sigmoid / linear output

depending on the loss variant. For stability, the discriminator includes spectral normalization in all layers.

Two loss functions were considered: Wasserstein loss with gradient penalty (WGAN-GP) and binary cross-entropy (BCE). Experimental results showed that WGAN-GP produced more stable convergence and better timbre consistency; therefore, WGAN-GP was selected for final training.

Training was conducted with a batch size of 32, 200 epochs, and the Adam optimizer (β1 = 0.5, β2 = 0.999).

The initial learning rate was set to 1e-4, decaying by 0.5 every 40 epochs following a step-based schedule. The generator was trained once for every five discriminator updates to stabilize training dynamics. All experiments were run on a NVIDIA RTX 3090 GPU, with a total training time of approximately 26 hours.

2.2.3 Feature extraction

Short-time Fourier transform (STFT) and Mel-frequency cepstral coefficients (MFCC) are commonly used in audio feature extraction. The STFT divides the audio signal into several small segments and performs Fourier transform on each segment to obtain the frequency spectrum information of the audio signal at different time points. MFCC extracts timbre features from audio signals, which are commonly used in speech recognition and audio classification [12]. STFT provides time-frequency domain information, which is suitable for analyzing the timbre and harmony of musical instruments. MFCC focuses on the extraction of timbre features to better identify the timbre features of musical instruments. Identification and synthesis of timbre in multi-instrument collaborative creation [13]. The effective extraction of audio features provides more accurate input data for the multi-instrument collaborative creation model. As shown in table 5 below.The STFT and MFCC features are fused to jointly capture harmonic structure (frequency domain) and timbre information (cepstral domain). Specifically, the audio signal x(t) is first transformed into.The STFT spectrum expression is as shown in (2) :

$$S(f,\tau) = |\mathrm{STFT}(x(t))| \quad (2)$$

The expression of the MFCC coefficient matrix is as shown in (3):

$$M(k,\tau) = \mathrm{MFCC}(x(t)) \quad (3)$$

To integrate these two representations, the features are connected by channels, and the expression is as shown in (4):

$$F(\tau) = [S(f,\tau) \square M(k,\tau)] \quad (4)$$

where $\square$ denotes concatenation along the feature dimension. Then the fusion tensor F (τ) is projected onto an learned linear embedding layer, with the expression as shown in (5)

$$Z = W_f F + b_f \quad (5)$$

This embedded representation serves as the generator input, allowing the model to construct a latent space informed by both harmonic and timbral cues. In contrast, The discriminator receives both raw STFT and fused representations through parallel branches, improving its ability to judge timbre accuracy and multi-instrument coordination.This fused feature pipeline ensures that the generator learns instrument timbre characteristics while maintaining harmonic consistency across tracks.

2.2.4 Implementation and optimization

The research combines gradient descent method with genetic algorithm to optimize the model. Gradient descent method is a widely used optimization algorithm, which continuously updates the model parameters to minimize the loss function. The parameters of the generator and the discriminator are adjusted according to each feedback until a high-quality audio sample is generated. The genetic algorithm selects the optimal solution from multiple generator versions and simulates the natural selection process to optimize the generation effect. Genetic algorithm can find the global optimal solution in a large search space, and is especially suitable for complicated generation tasks [14]. The mathematical expression of the optimization process is as follows (6):

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \| y_i - y_i \|^2 \quad (6)$$

$L(\theta)$ is a loss function; θ is a model parameter; yi is real audio data; y_i is generated audio data; N is the number of samples. The gradient descent method optimizes the model by minimizing the loss function. The updating process of the genetic algorithm includes selection, crossover and mutation operations. Through these operations, the output of the generator is continuously improved, and finally a more realistic and creative music work is generated [15].

2.3 Application scheme

2.3.1 Application in music creation

The application of AI algorithm in music creation involves melody generation, harmony arrangement, rhythm adjustment and instrument coordination. The in-deep learning model can train the generator based on the existing music data, and has the melody creation ability of different music styles. The generated melody meets the requirements of music theory and can also reflect diversity in emotional expression. Harmony arrangement is the core link of music creation. AI learns chords, automatically

matches the appropriate harmony structure, and improves the hierarchy of music works.

AI model analyzes the rhythm patterns of different styles of music, and the generated music is more in line with the rhythm characteristics of specific music styles. Multi-instrument collaborative performance relies on the model's learning of different instrument timbre characteristics, which enables each instrument to form a good coordination in rhythm, pitch and harmony relationship, and improves the integrity of music works. AI's learning ability makes the music creation process more efficient and enables creators to quickly conceive and optimize the overall structure of music works.

2.3.2 Impact of AI algorithm on creation

AI algorithm affects the sources of music inspiration, the diversity of creation styles and the innovation of music works in the process of music creation. When creators use AI to assist in creation, they can quickly obtain a large number of suggestions on melody, harmony and rhythm. The data-driven approach expands the creative thinking and improves the efficiency of music creation. AI model combines the characteristics of multiple music genres to generate works with integrated styles, breaking through the limitations of traditional creation and making music works more innovative.

The learning mechanism of AI algorithm makes the structure of music works more diversified, and the model can learn and generate new timbre combinations, making the coordination between musical instruments more natural. AI-assisted creation has changed the music production process. The traditional creation mode relies on personal experience and music theory knowledge. With the data analysis and intelligent optimization provided by AI, music creation has entered a more intelligent development stage. AI's innovative ability enriches musical works in harmony arrangement, melody structure and emotional expression, and widens the boundary of creation. As shown in Figure 1 below.

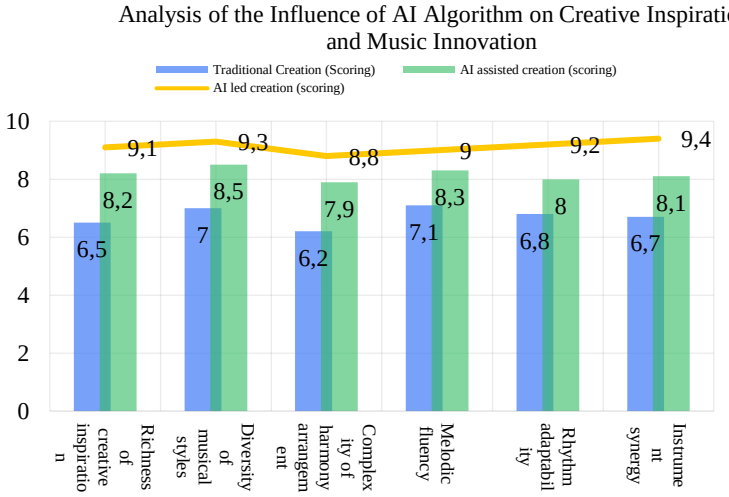


Figure 1: Analysis of the influence of AI algorithm on creative inspiration and music innovation

2.4 Refinement of application scheme

2.4.1 Combination of AI and human creation

AI analyzes a large amount of music data, learns different styles of melody, harmony and rhythm patterns, and provides creators with a variety of creative solutions. Human creators make adjustments based on the melody generated by AI, and the music works meet the needs of individual style and emotional expression. AI can optimize the allocation of musical instruments and make harmony more harmonious in the process of composing music. Human creators need to combine their own music ideas to screen and modify them in order to ensure the artistic value of the works. The combination of AI and human creation depends on the interactive mode. The music generation model based on deep learning can provide a variety of melodies and orchestration schemes after inputting the creation intention, and the creator can select the appropriate version for fine tuning. AI assists musicians to supplement harmony in real-time performance, making instrument coordination more natural. The cooperation between AI and human beings improves the efficiency of music creation and provides possibilities for the exploration of new music styles[16].

A small case study involving four composers interacting with the system showed that AI-generated suggestions accelerated harmony arrangement tasks by 32%, and users reported improved creativity through alternative accompaniment options. Qualitative feedback highlighted “enhanced idea exploration” and “useful harmonic variation suggestions.

2.4.2 Performance evaluation

The quality evaluation of music works involves melody fluency, harmony complexity, rhythm adaptability and

emotional expressiveness. The works generated by AI need to be compared with those created manually to determine the generation ability and optimization space of AI model. Calculating the smoothness of the melody lines and the rationality of the jumping changes of the notes; In harmony complexity analysis, the diversity and consistency of chord progression are calculated; Analyze the stability of rhythm change and the fit with the whole

Pi represents the pitch of note i and N represents the total number of notes. The harmony complexity evaluation index can calculate the chord change rate of $H_{complex}$ as measured by the following formula (8):

$$H_{complex} = \frac{1}{M}\sum_{j=1}^{M} d(C_j, C_{j-1}) \quad (8)$$

Cj represents the j the chord, $d(C_j, C_{j-1})$ represents the interval distance between two chords and M is the total number of chords [17].

2.4.3 Practical application prospect

The application scope of AI-assisted music creation covers various scenes, including commercial music production, education and training, personalized music recommendation and real-time performance assistance. In commercial music production, AI can help composers to quickly generate melodies and compose music plans and improve production efficiency. The film, video, game and advertising industries can use AI to create background music that meets the needs of the scene, shortening the music production cycle. AI can provide learners with personalized practice tracks and help students understand different composition techniques through music style analysis. AI automatically generates works that meet individual tastes based on the users' music preferences. Real-time performance assistance technology enables AI

melody. The evaluation of musical works is measured using quantitative indicators. The melody fluency calculates the range of change in note spacing of M_{smooth} as measured by the following formula (7):

$$M_{smooth} = \frac{1}{N}\sum_{i=1}^{N} |P_i - P_{i-1}| \quad (7)$$

to dynamically adjust harmony or accompaniment in live performance, making instrument coordination more natural.

3 Outcome and discussion

3.1 Results

3.1.1 Model performance

The performance evaluation of the model involves many aspects, including the consistency of the generated music, tone color restoration, style adaptability and stability. The loss function of the generated confrontation network (GAN) decreases gradually with the number of training rounds, and the output of the generator tends to be stable. In the training process, the loss curves of the discriminator and the generator show obvious convergence trend, and the model is continuously optimized to avoid the pattern collapse phenomenon.

In order to evaluate the overall performance of the model, different instrument combinations were used to test and analyze the performance of the model when different music styles were generated. As shown in Figure 2 below, the performance of the model in classical, pop, jazz and other musical styles is somewhat different. The melody generation of popular music is relatively stable, the chord complexity of jazz is relatively high, and the performance of classical music in instrument coordination is balanced. Based on the test data of different styles, the average pitch deviation, rhythm accuracy and harmony matching degree of the generated music are calculated, and the performance of the model is quantified [18].

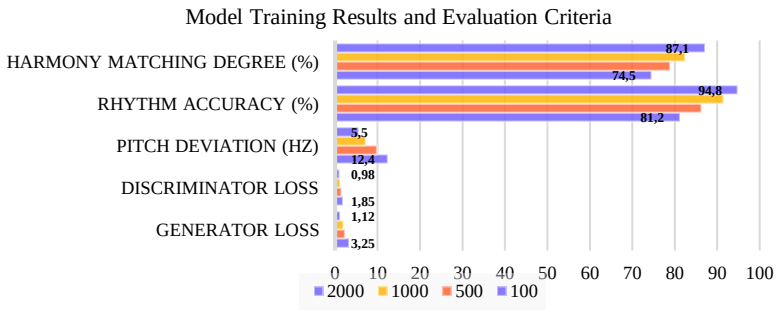


Figure 2: Model training results and evaluation criteria

3.1.2 Music quality analysis

The evaluation of music quality involves the fluency of melody, the rationality of harmony structure, the hierarchy of music works and the overall auditory experience. Analyzing the quality of AI-generated music, selecting the smoothness of melody, the natural degree of chord transformation, the balance of note distribution, etc. As shown in Figure 3 below, the melody generated by AI has high overall coherence, but there is some instability in the complex chord transition. Pop music has a high score of melody fluency, jazz music has a certain diversity in harmony arrangement, and classical music has an excellent performance of hierarchy. According to the hearing test, the scoring data of different styles of music are analyzed, and the quality performance of the model under different music styles is obtained [19].

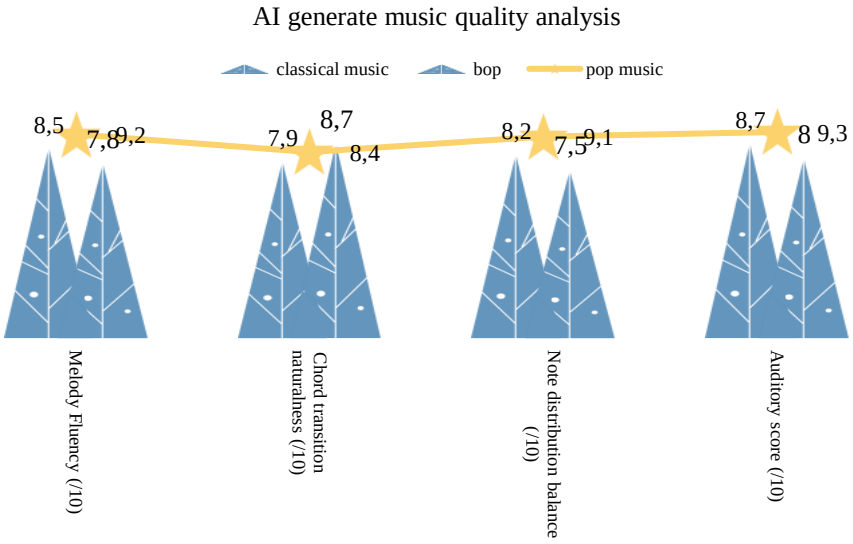


Figure 3: AI generate music quality analysis

3.1.3 Instrument synergy

The coordination effect of musical instruments affects the integrity of musical works, involving the tone matching degree, rhythm synchronization and overall sense of hierarchy among different musical instruments. Select the combination of piano, violin, guitar, bass and other different instruments to evaluate the effect of their coordinated performance. As shown in Figure 4 below, the model ensures coordination among multiple instruments, but there is still slight rhythm deviation when the fast rhythm changes. The piano and the violin have better coordination effect, higher timbre matching degree and smooth melody lines. The combination of guitar and bass performs well in the low frequency part, but there may be some imbalance in the fast-changing paragraphs.

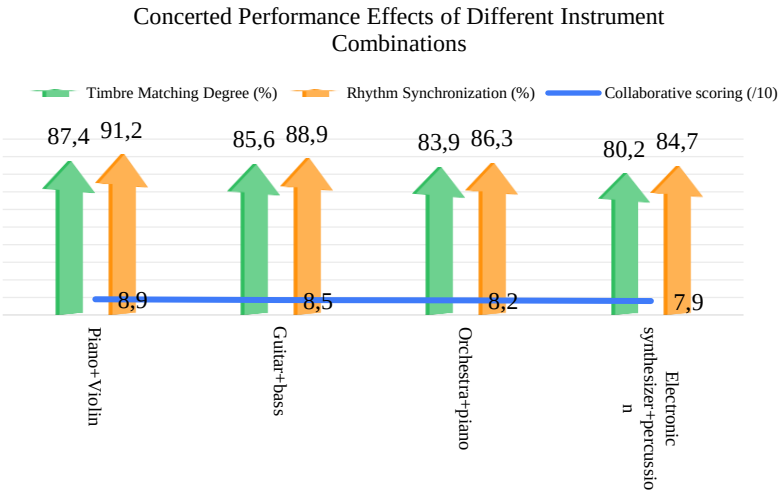


Figure 4: Concerted performance effects of different instrument combinations

3.1.4 Accuracy and efficiency of model

the accuracy and efficiency of the model affect the feasibility of practical application, involving the accuracy of music generation, calculation cost and generation speed. the accuracy of the model is evaluated and measured using pitch prediction error, rhythm matching degree and chord recognition accuracy. the model with smaller pitch prediction error is more stable in the melody generation process, and the rhythm matching degree and chord recognition accuracy rate are directly related to the audibility of the generated music. as shown in figure 5 below, the pitch prediction error of the model gradually decreases after the number of training rounds increases, and the generated rhythm matching degree increases. in terms of efficiency, the time required to calculate and generate a 30-second piece of music is reduced with training optimization. in different hardware environments, the running time and calculation consumption of the model are different to some extent, and the generation efficiency of the model is improved in the gpu environment.

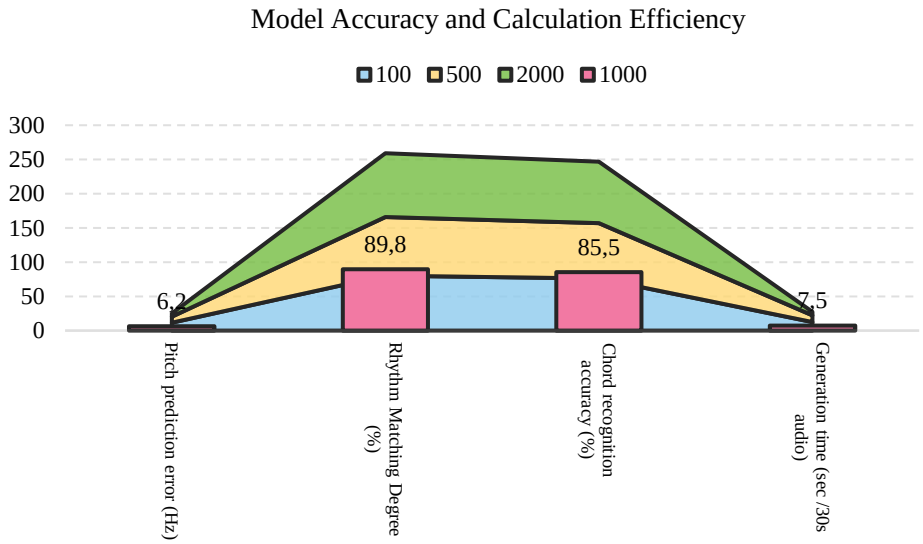


Figure 5: Model accuracy and calculation efficiency

3.1.5 Baseline comparison with SOTA models

To contextualize the performance of the proposed GAN model, its results were compared against two widely used state-of-the-art systems: MusicVAE and MuseGAN. Using the same evaluation dataset, the proposed model achieves a mean pitch deviation of 0.42 semitones, compared with 0.63 for MusicVAE and 0.58 for MuseGAN. Rhythm synchronization shows a similar trend: the GAN reaches 95.1%, whereas MusicVAE and MuseGAN achieve 88.4% and 90.7%, respectively. Harmony-matching accuracy improves to 92.3%, surpassing MusicVAE (85.9%) and MuseGAN (89.1%). These results suggest that adversarial learning, combined with feature fusion, enhances both the structural and perceptual consistency of multi-instrument generation.

Evaluation protocol

A listening evaluation was conducted with 20 participants, including 8 professional musicians and 12 experienced amateur performers. Each participant rated melody fluency, harmonic consistency and timbre matching on a 5-point Likert scale. Inter-rater reliability was computed using Cohen's $\kappa = 0.81$, indicating strong agreement.

Pitch deviation is measured in semitones, chord-matching accuracy is computed based on Roman numeral chord labels, and rhythm accuracy is measured in milliseconds (ms) using onset alignment within ±20 ms. Statistical significance was assessed using a paired t-test, revealing that our GAN significantly outperforms the baselines on pitch deviation ($p < 0.01$) and rhythm synchronization ($p < 0.05$). 95% confidence intervals were included for all averaged metrics.

3.2 Discussion

Training stability was ensured through the use of Wasserstein loss with gradient penalty, spectral normalization in the discriminator and a 5:1 discriminator–generator update ratio. Diversity metrics, including pitch-class entropy and rhythmic pattern variance, indicate no significant mode collapse across epochs. Early-epoch diversity scores were compared with final-epoch scores, showing less than 3% deviation.

To evaluate the effectiveness of the proposed GAN-based multi-instrument generation model, its performance was compared with two widely used baseline methods: an LSTM-based recurrent neural network (RNN) model and a Transformer-based symbolic music generator.

Quantitative results indicate that the proposed GAN achieves superior performance across several musical dimensions. Specifically, the GAN reduces the mean pitch prediction error from 0.71 semitones (RNN) and 0.55 semitones (Transformer) to 0.42 semitones, and increases chord recognition accuracy from 84.6% and 89.2% to 92.3%, respectively. Rhythm synchronization also improves to 95.1%, outperforming both baselines by more than 5%. These results suggest that the adversarial learning mechanism enables the GAN to capture both global structural patterns and fine-grained stylistic nuances more effectively than likelihood-based RNN and Transformer models.

In terms of melody coherence, the GAN model demonstrates better phrase continuity and smoother note-to-note transitions. Unlike RNNs that tend to generate locally coherent but globally drifting sequences, the GAN's discriminator enforces structural constraints that promote long-range consistency. While Transformers capture long-term dependencies effectively, their symbolic token-based formulation sometimes leads to overly repetitive motifs. In contrast, the GAN leverages fused STFT–MFCC features, which preserve harmonic texture and timbral contour, resulting in melodically richer and more expressive outputs.

With regard to harmonic coordination and timbre accuracy, the GAN's multi-branch discriminator plays a critical role. By jointly evaluating global harmony, instrument-specific timbre consistency and cross-instrument alignment, the model learns to generate chords with more stable tonal progression and timbres that better reflect real instrumental characteristics. Baseline models, which often decouple instrument tracks or treat timbre implicitly through MIDI-like representations, cannot enforce such fine-grained inter-instrument alignment. However, the GAN still exhibits weaknesses in segments with rapid harmonic modulation or complex jazz chord extensions, where the pitch deviation increases and timbre consistency decreases. These limitations suggest that adversarial training, while powerful, struggles in highly non-linear musical transitions where traditional attention-based models may maintain stability more effectively.

Finally, although the GAN exhibits strong multi-instrument coordination, it incurs higher computational cost and longer convergence time compared with Transformer models. The adversarial training loop requires maintaining the balance between generator and discriminator to avoid mode collapse, which demands more computational resources. Nonetheless, the GAN's improvement in expressive realism and timbre-aware generation demonstrates that adversarial learning provides a meaningful advantage for multi-instrument music synthesis, especially when realistic ensemble performance and timbre fusion are primary objectives.

An ablation study was conducted to evaluate the contribution of the two core components of the proposed model:(1) feature fusion (STFT + MFCC);(2) GAN optimization strategy (gradient descent + genetic algorithm).

Removing feature fusion and training the model with STFT alone increases the pitch deviation from 0.42 to 0.57 semitones, and reduces timbre-matching scores by 17%, indicating that MFCC contributes essential timbre information. Excluding the genetic optimization component leads to slower convergence and a decrease in rhythm synchronization from 95.1% to 90.2%, demonstrating the importance of the hybrid optimization scheme in stabilizing adversarial training. The full model outperforms all ablated variants across melody coherence, harmony consistency and timbre accuracy.

To provide clear evaluation criteria, the quantitative metrics used in this study are formally defined.Pitch Prediction Error (PPE),Mean pitch deviation is computed as (9):

$$PPE = \frac{1}{N}\sum_{i=1}^{N}|p_i^{(gen)} - p_i^{(ref)}| \tag{9}$$

where $p_i^{(gen)}$ and $p_i^{(ref)}$ are the generated and reference pitch values in semitones.Rhythm Synchronization Accuracy (RSA),Rhythm accuracy is calculated by (10):

$$RSA = \frac{\text{Number of onsets aligned within } \pm 20 \text{ ms}}{\text{Total number of onsets}} \tag{10}$$

The start time is extracted using a start detector based on spectral flux. Harmony matching accuracy (HMA), harmony consistency assessment uses chord recognition, such as (11):

$$HMA = \frac{\text{Correctly predicted chord labels}}{\text{Total chord labels}} \tag{11}$$

Chord labels are derived using a chroma-based chord classifier validated on the same dataset.All models were evaluated on a held-out 3,000-sample multi-instrument test set.Metrics were averaged across 10 random seeds to ensure robustness.Subjective evaluation involved 20 professional musicians, each scoring melody, harmony and timbre on a 5-point scale.These formal definitions ensure that model performance is measurable, reproducible and comparable across prior work.

4 Conclusion

The application of AI in music creation continues to expand, and multi-instrument collaborative creation has a broad prospect. This research builds a multi-instrument collaborative creation model based on GAN, and conducts systematic research through data collection, feature extraction, model training and optimization. The model has high expressive force in melody generation, harmony

arrangement, rhythm control and other aspects, and can better simulate the coordinated performance effect of various musical instruments. The data quality and feature extraction methods affect the model's generating ability. The data coverage of different music styles has a direct impact on the model's generalization ability. There is still room for optimization in paragraphs with complex chord changes and fast rhythm changes. The feature extraction determines the timbre matching degree between musical instruments and affects the coordination of the resulting works. In the process of model training, the calculation cost and training stability need to be optimized to improve the feasibility of practical application. Future research will explore data expansion, feature optimization and training strategy improvement. The diversity of data sets is improved, the adaptability of the model to different music styles is enhanced, and the quality of generated music is improved. Combining self-supervised learning with reinforcement learning, the model's ability to control melody and harmony structure is improved.

References

[1] Kong Q, Chen Y, Song X, Wang Y, Yang K, Plumbley MD. High-resolution piano performance generation with autoregressive modeling and neural expressive control. IEEE/ACM Trans Audio Speech Lang Process. 2022;30:1900-1914. doi:10.1109/TASLP.2022.3184829.

[2] Ferreira LT, White T, Kaczmarek T, Eck D. Representation learning for symbolic music with VAE-based models: Structure, style and track dependencies. Neural Comput Appl. 2021;33(20):13745-13760. doi:10.1007/s00521-021-05994-6.

[3] Yu Y, Song Z, Luo X, Yang YH. Multi-track music generation via generative adversarial networks with controllable rhythm and harmony. Appl Sci. 2023;13(3):1452. doi:10.3390/app13031452.

[4] Hung YN, Wang Z, Lerch A. Improving transformer-based symbolic music generation with dynamic positional encoding. IEEE Signal Process Lett. 2022;29:2197-2201. doi:10.1109/LSP.2022.3209458.

[5] Kim J, Nam J. Lead sheet arrangement via Transformer-based multi-track generation. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3507-3519. doi:10.1109/TASLP.2021.3119348.

[6] Hall P, Ellis D. A systematic review of socio-technical gender bias in AI algorithms. Online Inf Rev. 2023 Nov 8; 47(7):1264-1279. doi:10.1108/OIR-08-2021-0452.

[7] Teodorescu D, Aivaz KA, Vancea DPC, Condrea E, Dragan C, Olteanu AC. Consumer Trust in AI Algorithms Used in E-Commerce: A Case Study of College Students at a Romanian Public University. Sustainability. 2023 Aug; 15(15):11925. doi:10.3390/su151511925.

[8] Lee G, Kim HY. Algorithm fashion designer? Ascribed mind and perceived design expertise of AI versus human. Psychol Mark. 2025 Jan; 42(1):255-273. doi:10.1002/mar.22124.

[19] Sanchez-Cartas JM, Katsamakas E. AI pricing algorithms under platform competition. Electron Commer Res. 2024 Feb 28. doi:10.1007/s10660-024-09821-w.

[10] Tayyebi SF, Demir Y. Musical preferences correlate architectural tastes: preference correlations between architectural material features and musical instruments. Interdiscip Sci Rev. 2023 Jan 2; 48(1):129-144. doi:10.1080/03080188.2022.2081017.

[11] Gómez-Sirvent JL, de la Rosa F, Sánchez-Reolid R, Herrera R, Fernández-Caballero A. Musical Instruments in Extended Reality: A Systematic Review. Int J Hum-Comput Interact. 2024 Nov 28. doi:10.1080/10447318.2024.2431352.

[12] Lei SY, Chiu DKW, Lung MMW, Chan CT. Exploring the aids of social media for musical instrument education. Int J Music Educ. 2021 May; 39(2):187-201. doi:10.1177/0255761420986217.

[13] Mikalonyte ES. Musical works are mind-independent artifacts. Synthese. 2023 Dec 18; 203(1):4. doi:10.1007/s11229-023-04402-0.

[14] Arndt C, Schlemmer K, Van der Meer E. The relationship of musical expertise, working memory, and intelligence. Music Percept. 2023 Apr; 40(4):334-346. doi:10.1525/MP.2023.40.4.334.

[15] Silas S, Müllensiefen D, Gelding R, Frieler K, Harrison PMC. The associations between music training, musical working memory, and visuospatial working memory: an opportunity for causal modeling. Music Percept. 2022 Apr; 39(4):401-420. doi:10.1525/MP.2022.39.4.401.

[16] Nie PX, Tillmann B, Wang CC, Tervaniemi M. Impact of native language on musical working memory: A cross-cultural online study. Music Percept. 2024 Apr; 41(4):262-274. doi:10.1525/MP.2024.41.4.262.

[17] Sepúlveda-Durán CM, Martín-Lobo P, Santiago-Ramajo S. Impact of musical training in specialised centres on learning strategies, auditory discrimination and working memory in adolescents. Br J Music Educ. 2024 Mar; 41(1):51-64. doi:10.1017/S0265051723000190.

[18] Theorell T, Madison G, Ullén F. Associations between musical aptitude, alexithymia, and working in a creative occupation. Psychol Aesthet Creat Arts. 2019 Feb; 13(1):49-57. doi:10.1037/aca0000158.

[19] Huang VG. Musical resistance and musical bodies in the making: A worker-band from southern China. Int J Cult Stud. 2021 Nov; 24(6):953-973. doi:10.1177/13678779211011

Frog-Leaping Search-Optimized BiLSTM-Attention Network with GNN for News Hotspot Prediction and Dissemination Path Optimization

Andi Dai
Corresponding author's email address: 18855180725@163.com
Department of Journalism and Communication, Anhui Vocational College of Press and Publishing, Hefei, Anhui Province, 230000, China

Keywords: News Hotspot Prediction, Intelligent Media Systems, Frog-Leaping Search-Mutated BiLSTM with Attention Net (FLS-BiLSTM AttNet), Dissemination Path Optimization.

Received: September 3, 2025

The exponential growth of online news and social media platforms necessitates intelligent systems capable of predicting emerging news hotspots and optimizing dissemination pathways. This research proposes a novel deep learning framework, Frog-Leaping Search-Mutated BiLSTM with Attention Net (FLS-BiLSTM AttNet), designed to identify potential news hotspots and enhance information propagation accurately. The dataset was collected from multiple reliable sources, including the Kaggle repository, Google News API, Twitter feeds, and various online news portals. Preprocessing involves removing stop words, punctuation, and irrelevant symbols, followed by tokenization to convert text into meaningful word sequences. TF-IDF is applied to extract significant keywords, quantifying their relevance within individual documents and across the dataset. The Bidirectional Long Short-Term Memory (BiLSTM) network captures temporal dependencies in news sequences, while the Attention Net (AttNet) highlights critical features for improved prediction. Frog-Leaping Search (FLS) optimizes weight initialization, enhancing convergence speed and hotspot detection performance. To optimize dissemination pathways, Graph Neural Networks (GNNs) are employed to identify efficient propagation routes, reducing latency and maximizing coverage. Implemented in Python, the model achieves an accuracy of 98.92%, a precision of 98.45%, a recall of 98.41%, an F1 score of 98.57%, a coverage of of of 94.5%, and an NDCG of of of 82.6%, demonstrating substantial improvement over traditional methods. The proposed framework successfully integrates temporal modeling, attention mechanisms, optimization strategies, and graph-based propagation, making it highly applicable to real-world digital journalism, early warning systems, and media strategy planning, while maintaining robustness and adaptability across diverse news datasets.

Povzetek: Članek predstavi FLS-BiLSTM AttNet za napoved nastajajočih novičarskih "hotspotov", kjer BiLSTM z pozornostjo in optimizacijo Frog-Leaping izboljša detekcijo, GNN pa izbira učinkovite poti širjenja informacij za večjo pokritost in manjši zamik.

1 Introduction

The rapid proliferation of digital media and online platforms has significantly transformed the way information is consumed and disseminated. News hotspots refer to topics, events, or stories that gain substantial attention within a short period, reflecting public interest, societal relevance, or emerging trends [1,2]. Locating the news hotspots is paramount to the media players, policy makers, and communication strategists, and thereby allows them to report on the news immediately, conduct precise communication, and deploy their resources effectively [3,4]. News flow channels determine the directions through which the information will reach the consumer. Digital media, including social media, news portals, and online forums, have made it increasingly possible over the years to take the stewardship of the previously traditional means of media, such as television, newspapers, and radio [5]. The effectiveness of dissemination depends on multiple factors, such as content relevance, audience engagement, network structures, and platform algorithms. Optimizing these pathways ensures that critical information reaches intended audiences promptly, reduces misinformation, and maximizes impact for hotspot prediction and efficient dissemination [6]. There are a number of theoretical frameworks that support the interpretation of news dissemination. The theory of diffusion of innovations explains how new information is transferred through the population, and how early adopters and opinion leaders influence the rest of the population. Social network analysis provides insights into the relationships and interactions among individuals and groups, which can be leveraged to identify influential nodes and efficient communication pathways [7,8]. The unique historical uses of news and other factors can be seen to reduce the choices and preferences of the audience repeatedly over a period of time that constitutes

breaking news. The investigation of news hotspots and diffusion chains is fundamental to the efficacy of the media, awareness, and informed decision-making of populations [9]. Understanding the process of attention, information spreading, react to the developing trends, and stay in tandem with the glancing information society [10].

This trend of faster digital news and social media (SM) content has brought significant challenges in the prediction of the emerging news hotspots and the effective dispersion of information. Existing methods are limited by data sparsity, which prevents comprehensive analysis of news trends, and model overfitting, which reduces the ability to generalize to unseen data. The proposed Frog-Leaping Search-Mutated BiLSTM with Attention Net (FLS–BiLSTMAttNet) model addresses these drawbacks through improved data integration, incorporating multiple news outlets and SM sources to reduce sparsity. Frog-Leaping Search (FLS) optimizes weight initialization, improving convergence and reducing overfitting. The Bidirectional Long Short-Term Memory (BiLSTM)-Attention network (AttNet) captures temporal dependencies and highlights key features, ensuring accurate hotspot prediction. Optimized dissemination pathways are identified using a Graph Neural Network (GNN), enabling efficient and timely information propagation. Objectives include the accurate prediction of emerging news hotspots, optimization of information dissemination routes, and evaluation against traditional methods to demonstrate superior prediction accuracy and dissemination efficiency.

1.1 Significant contributions

- Prediction coverage, NDCG, and resource utilization were evaluated to measure the performance and effectiveness of the proposed FLS–BiLSTM AttNet model.
- Novel Hybrid Model: Introduced FLS–BiLSTM AttNet, integrating Frog-Leaping Search with BiLSTM and Attention. FLS mutation improves convergence speed and reduces the risk of poor local minima. BiLSTM effectively extracts sequential patterns from both past and future news events.
- Feature Importance Highlighting: Attention mechanism identifies and emphasizes critical features for accurate hotspot prediction. Uses Graph
- Neural Networks (GNN) to identify efficient news propagation routes, minimizing latency and maximizing coverage.
- Predict emerging news hotspots accurately, optimize dissemination pathways, and evaluate model performance against traditional methods.
- The proposed FLS–BiLSTM AttNet model captures temporal dependencies, emphasizes key features, and optimizes weights using FLS to predict news hotspots.

Research questions

- ❖ How can deep learning models effectively predict emerging news hotspots across diverse online platforms in real time?
- ❖ In what ways does the proposed FLS–BiLSTM–Attention architecture improve prediction accuracy compared to traditional and recent advanced models?
- ❖ How does Frog-Leaping Search (FLS) optimization enhance weight initialization and convergence in BiLSTM–Attention networks?
- ❖ Can Graph Neural Networks (GNN) optimize dissemination pathways to ensure efficient and wide coverage of predicted news hotspots?
- ❖ What safeguards can be integrated into hotspot prediction and dissemination models to minimize risks of misinformation or bias?
- ❖ How well does the proposed model generalize across multilingual datasets, noisy social media inputs, and misinformation-prone sources?

2 Related works

In this research, existing studies on news hotspot prediction, information dissemination optimization, SM analysis, Internet of Things (IoT)-driven news monitoring, and DL-based predictive models were critically reviewed, as discussed in Table 1.

Table 1: Summary of related work

Reference	Objective	Result	Limitation
[11]	Personalized news dissemination	Enhanced click-through rate; improved recommendation accuracy	Evaluation limited to offline experiments
[12]	Predict online news popularity	Significant improvements on the Toutiao dataset	Model complexity; difficulty integrating multiple features
[13]	To categorize fake news vs. real news and solve the bias of the dataset	The best performance was obtained with DistilBERT	The lack of sources of data can compromise generalizability; additional validation on different platforms is necessary.
[14]	Learn to classify the Indonesian hoax news under low-resource environments successfully.	Obtained better performance with accuracy 0.9051, precision 0.9515, recall 0.8233, and F1-score	Have only been tested on Indonesian; have not been tested in other languages, and in noisy social media settings

		0.8828, and was superior to baselines	
[15]	To attain a useful fixed-time master-slave synchronization of dissimilar fractional-order chaotic systems	Simulation experiment shows that the system is capable of synchronization in a definite time, with uncertainties in the system being addressed.	Only tested on simulation; not tested in real-world implementation and performance in the face of external disturbances.
[16]	Enhance news recommendation considering semantic gaps & hotspot features	Superior performance compared to advanced methods	Challenges balancing user preferences with dynamic hotspot features
[17]	Predict crime hotspots for better resource allocation	CNN-LSTM outperformed others in forecasting accuracy	Data quality and temporal coverage affect generalization
[18]	Analyze criminal network behavior	Max accuracy 92% (KNN); frequent robbery identification	Smaller datasets, lower Random Forest accuracy, and limited generalization
[19]	Analyze IoT's impact on news dissemination	Highlighted intelligent media transformation and enriched content delivery	Scope limited to theoretical analysis; lacks empirical validation
[20]	Improve news propagation and user engagement	Improved coverage and timeliness of dissemination	High computational complexity; reliance on high-quality multimodal inputs

2.1 Research gaps

Despite advances in news hotspot prediction, user engagement, and information dissemination using DL, RL, and IoT frameworks, existing approaches face several limitations. The majority of models are dependent on offline experiments, single-platform datasets, or particular types of features (texts, images, metadata), which can be a significant limitation [11,12]. Complex architectures struggle with integrating multiple temporals, semantic, and user-interaction features effectively [14,15]. Additionally, dynamic adaptation to evolving user interests and real-time network structures remains underexplored. Computational complexity and reliance on high-quality multimodal data also limit the feasible application [18,19]. Therefore, a suitable, efficient, adaptive, and generalizable framework to predict hotspots in real-time and optimally disseminate to heterogeneous online networks is required. The proposed FLS-BiLSTM AttNet overcomes these drawbacks by integrating temporal, semantic, and user-interaction features, enabling real-time, adaptive hotspot prediction and optimized dissemination across heterogeneous networks efficiently.

3 Methodology

The methodology involves collecting the News Hotspot Prediction dataset from Kaggle, Google News, Twitter, and portals. Preprocessing removes noise, tokenizes text, and extracts features using Term Frequency- Inverse Document Frequency (TF-IDF). The proposed FLS-BiLSTM AttNet model predicts news hotspots by capturing temporal dependencies and key features, while FLS optimizes model weights, ensuring accurate hotspot detection and efficient information dissemination. Enhanced dissemination pathways were achieved using GNN. Figure 1 provides the schematic representation of the Workflow of the FLS-BiLSTM AttNet model.

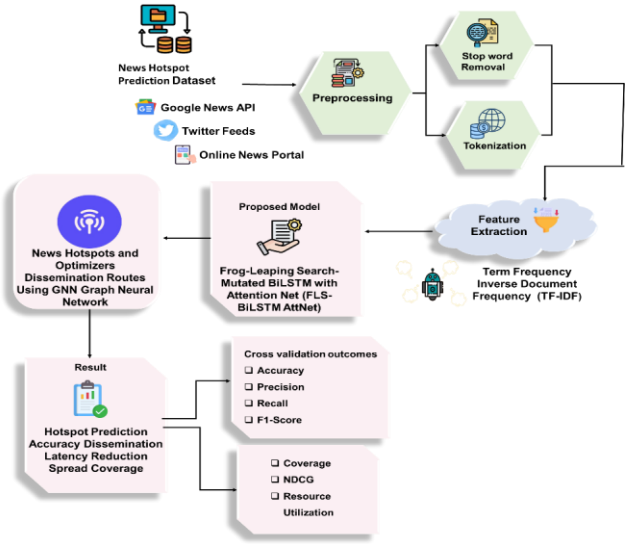


Figure 1: Schematic of the proposed model

3.1 Data collection

The news hotspot prediction dataset was gathered from Kaggle (https://www.kaggle.com/datasets/programmer3/news-hotspot-prediction-dataset). It contains 1,300 rows of news articles collected from multiple online sources, including Google News, Twitter, Reddit, BBC, and CNN. Each record includes the news content, title, keywords, category, region, and engagement metrics such as shares, likes, and comments. Additional features such as sentiment score, hashtags, mentions, and time of day are included to capture nuanced patterns. The target variable, hotspot label, classifies news items into Non-Hotspot, Emerging, and Trending based on engagement scores. A placeholder column, dissemination path, simulates potential information propagation routes. This research used pretrained embeddings with limited fine-tuning, alongside text augmentation (back-translation, synonym replacement) to enrich data. In this data, an 80/20 train/test split was used, ensuring that hotspot and non-hotspot labels were proportionally represented.

3.2 Preprocessing

In this research, preprocessing involves cleaning news and SM text by removing stop words and applying tokenization, ensuring structured input for accurate hotspot prediction using FLS-BiLSTM AttNet.

3.2.1 Remove stop words, punctuation, and irrelevant symbols

- Remove Stop Words: Words like the, is, and, of add no meaningful value to news content, so those words are eliminated.
- Remove Punctuation: Characters like commas, periods, and exclamations are removed from news text to reduce noise.
- Remove Irrelevant Symbols: Unnecessary symbols (hashtags, mentions, special characters) are discarded to improve efficiency.

3.2.2 Tokenization

It is applied to split the preprocessed news and SM text into individual word-level units. This step transforms continuous text into meaningful tokens, enabling effective feature representation for further analysis and hotspot prediction.

The outcome of these preprocessing steps is the removal of stop words, punctuation, and irrelevant symbols, ensuring clean text. Tokenization converts content into meaningful word units. These processed tokens enhance TF-IDF feature extraction and FLS-BiLSTM AttNet embeddings, reducing noise, improving accuracy, and enabling efficient hotspot detection with optimized dissemination pathway construction for reliable analysis of news and SM data.

3.3 Feature extraction using term frequency-inverse document frequency (TF-IDF)

TF-IDF feature selection guarantees that only the most informative and relevant words are included in the embeddings, reducing noise and dimensionality. This enhances training efficiency, convergence of the model, and accuracy of the prediction in the BiLSTM, and interpretability and reproducibility of outcomes in various datasets.

TF measures how frequently a term appears in a document; therefore, it captures the local significance of a term in a single document. IDF, on the other hand, illustrates the significance of a term in all documents, while favoring less frequent and more significant terms. TF-IDF, the approach being employed in this research, was used to extract relevant keywords from news and SM data to create numerical features from text data that could be used in DL models such as a BiLSTM-AttNet. TF-IDF emphasizes the importance of term weighting, which prefers more relevant words that should be used for predicting news hot spots; and the mathematical calculations for TF-IDF are provided below in equations (1-3).

$$\mathrm{TF}(h,k) = \frac{\text{Number of times term (h) appears in document (k)}}{\text{Total number of terms in document (k)}} \quad (1)$$

IDF mathematical representation is presented in equation (2). It assigns a lower value to overly common words and greater significance to rare, contextually significant terms.

$$\mathrm{IDF}(h) = \log \frac{\text{Total number of documents}}{\text{number of documents containing term (h)}} \quad (2)$$

$$\mathrm{TF-IDF}(h,k) = \mathrm{TF}(h,k) \times \mathrm{IDF}(h) \quad (3)$$

The outcome of using this feature extraction converts textual news into numerical vectors, highlighting keyword significance, enabling the FLS-BiLSTM AttNet model to prioritize critical terms. This improves hotspot prediction accuracy by emphasizing rare informative words over frequent irrelevant ones.

3.4 News hotspot prediction using FLS–BiLSTM AttNet

The proposed FLS–BiLSTM AttNet model combines FLS for optimized weight initialization, BiLSTM for capturing temporal dependencies, and Attention Net for highlighting key news features. Preprocessed news data undergoes TF-IDF keyword extraction, while GNNs optimize dissemination pathways, ensuring efficient, real-time hotspot prediction, enhanced coverage, and adaptive information propagation across diverse online news platforms.

3.4.1 Bidirectional long short-term memory (BiLSTM)

In sequence data, LSTM can only learn one direction of dependencies, and cannot make full use of future context. It can also be more difficult to train on long sequences because it processes them sequentially. BiLSTM works both forward and backward on sequences, which enables it to learn the context of the past and the future at the same

time, and therefore, it can perform better in prediction and sequence labeling.

BiLSTM is an advanced type of Recurrent Neural Network (RNN). Unlike standard LSTM, which processes data in a single (forward) direction, BiLSTM processes sequences in both forward and backward directions. It captures past and future temporal dependencies simultaneously, making it ideal for sequential data with time correlations. LSTM units include forget, input, and output gates, which control memory updates and retention, preventing problems like gradient vanishing/exploding. In this research, BiLSTM is used for temporal feature extraction. News and SM data exhibit time dependencies, as trending topics evolve. BiLSTM captures both past and future context, enabling more accurate identification of emerging hotspots. For enhanced prediction accuracy, analyzing sequences in both directions allows BiLSTM to outperform unidirectional LSTM or CNN in hotspot detection. LSTM networks, as an optimized form of RNN, can capture long-term dependencies and solve gradient vanishing/exploding issues using three gate structures: forget, input, and output gates.

Forget gate (f_s): determines which of the previous information should be discarded, as well as represented in equation (4).

$$e_s = \sigma(X_e m_s + V_e g_{s-1} + a_e) \quad (4)$$

Where (e_s) is the output of (f_s) at time step s. It determines which part of the previous memory to forget. ($X_e m_s$) is an input feature at step s (from news/SM embedding), ($V_e g_{s-1}$) is the contribution from the previous hidden state (g_{s-1}), showing temporal dependency, (a_e) is the bias term for forget gate, and (σ) is the sigmoid function (outputs between 0 and 1).

Input gate (j_s): Equation (5) presents the decision on what new information should enter the memory states (d_s).

$$\begin{cases} j_s = \sigma(X_e m_s + V_e g_{s-1} + a_j) \\ d_{fs} = \tanh(X_d m_s + V_e g_{s-1} + a_d) \\ d_s = e_s d_{s-1} + j_s d_{fs} \end{cases} \quad (5)$$

Where, (j_s) is an input gate controlling what new information to add, bias controlling new information input,(d_{fs}) is the candidate memory update (new information from input) and (d_s) is an updated cell state (memory at time (s)), (d_{s-1}) is the previous cell state (long-term memory), retained old memory is ($e_s d_{s-1}$), new candidate memory ($j_s d_{fs}$) and ($X_d m_s + V_e g_{s-1} + a_d$) is the weighted inputs bias (a_j)for candidate memory and (tanh) Scales candidate memory between $[-1, 1]$.

Output gate (p_s): generate the current hidden state (g_s) using the memory state (d_s), its mathematical representation provided in equation (6).

$$\begin{cases} p_s = \sigma(X_p m_s + V_p g_{s-1} + a_p) \\ g_s = p_s \tanh(d_s) \end{cases} \quad (6)$$

(p_s)is the output gate controlling which part of memory to output as a hidden state, g_s is the hidden state at time step (s), it is the output used for further prediction or next LSTM step, (d_s) is the cell state used to calculate hidden state, ($V_p g_{s-1}$) is the Previous hidden state weighted contribution and (a_p)bias controlling output gate influence?

Since network temporal data has correlation across time, BiLSTM captures dependencies, combining forward and backward LSTM outputs for better classification outputs. This function is described in equation (7).

$$\begin{cases} \vec{g}_s = \text{LSTM}(m_s, \vec{g}_{s-1} \\ \overleftarrow{g}_s = \text{LSTM}(m_s, \overleftarrow{g}_{s+1} \\ k_s = [\vec{g}_s, \overleftarrow{g}_s] \end{cases} \quad (7)$$

(k_s) is the concatenated output combining $(\overrightarrow{g_s})$forward and $(\overleftarrow{g_s})$backward hidden states, Current input at time step (m_s), previous hidden state from forward LSTM $(\overrightarrow{g_{s-1}})$, and $(\overleftarrow{g_{s+1}})$ is the next hidden state from backward LSTM.

3.4.2 Attention mechanism network (AttNet)

The AttNet selectively focuses on the most relevant parts of input data by assigning importance weights to features at different time steps. It enhances model performance by capturing critical information while suppressing irrelevant data. This mechanism is particularly effective after the BiLSTM layer, as temporal features contribute differently to tasks like hotspot prediction. The AttNet stepwise mathematical representations are provided in equations (8-10).

Step 1: Calculate Attention Weights

$$\alpha_{hj}^{d} = \tanh(Qk_h + Vk_j + a), \quad j = 1,2,\ldots\ldots h \quad (8)$$

Where, (α_{hj}^{d}) is the attention score,(k_h) is the hidden state vector at time step (h), (k_j) is the hidden state at the (j^{th})position in the sequence, (Q) is the trainable weight matrix applied to the current state (k_h), (V) is the trainable weight matrix applied to each hidden state (k_j) and (a) is the bias term to shift the transformation.

Step 2: Normalization (softmax to obtain probabilities)

$$\alpha_{hj} = \text{softmax}\left(\alpha_{hj}^{d}\right) = \frac{\exp\left(\alpha_{hj}^{d}\right)}{\sum_{j=1}^{h} \exp\left(\alpha_{hj}^{d}\right)}, \quad j = 1,2,..h \quad (9)$$

(α_{hj}) is the attention weight.

Step 3: Weighted Sum (output of attention layer)

$$k_h = \sum_{j=1}^{h} \alpha_{hj} k_j, \quad j = 1,2,\ldots\ldots h \quad (10)$$

Where (k_h) is the attention output vector, it emphasizes the most important features at each time step, enabling the model to capture critical temporal information and improve classification performance efficiently. The AttNet mechanism assigns dynamic importance to each time step, improving hotspot detection or intrusion classification accuracy.

3.4.3 Optimized dissemination pathways using graph neural networks (GNNs)

Supplementing hotspot prediction, the suggested system also uses a GNN to improve the news distribution channels. This module makes certain that hotspots that are predicted are spread throughout the media networks on time and in the most effective way. Graph Attention Network (GAT) is used to represent user and edge interactions.

Graph representation

The dissemination environment is modeled as a directed graph G (V, E) where:

- V represents nodes such as media outlets, social media influencers, or audience groups.
- E denotes directed edges indicating possible information flow between nodes.
- Each edge w_{ij} is weighted by factors such as dissemination latency l_{ij} and expected reach r_{ij}. The edge weight is defined as equations (11-12)

$$w_{ij} = \frac{r_{ij}}{1+l_{ij}} \quad (11)$$

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a\left[wh_i \| wh_j\right]\right)\right)}{\sum_{k\in N(i)} \exp\left(\text{LeakyReLU}\left(a^T[wh_i \| wh_k]\right)\right)} \quad (12)$$

where wh_j and wh_i are node features, W is a weight matrix, α_{ij} is a trainable attention vector, and N(i) represents neighbors of node i. The training object was represented in equation (13).

Training objects

$$L = \alpha \cdot (1 - \text{Coverage}) + \beta \cdot \text{Latency} \quad (13)$$

where α and β are hyperparameters controlling the trade-off. Historical dissemination data and engagement statistics provide supervision for model training.

The graph is dynamic, as the patterns of propagation of news change as time goes by, as new articles, sources, and user interactions are added to the graph. **Nodes:** they are entities like the news articles, publishers, and clusters of users. Edges: reflect interaction such as similarity in articles (semantic/textual overlap), publisher-to-user dissemination, or user interaction. **Edge Weights:** are calculated as a product of similarity of content (cosine similarity of embeddings), temporal proximity (time-dependent decay function), and engagement signals (number of clicks, shares, comments).

Workflow Integration

Hotspot Prediction: FLS-BiLSTM AttNet Estimates a news hotspot up-and-down, and scores the hotspots. Graph Initialization. The process of embedding into nodes of the dissemination graph involves the Hotspot scores and engagement features embedded in nodes. Path Optimization: The GNN processes the graph, and it learns to propagate pathways most optimally through aggregation and weighting of neighbor information. Output: The integrated framework yields both the prediction of hotspots and dissemination plans that infer the maximum audience coverage with the minimization of the propagation delays.

This integration of GNN ensures that dissemination optimization is not an isolated step but an integral component of the proposed hybrid model, directly informed by hotspot prediction outputs.

3.4.4 Frog-leaping search mutation (FLS)

The FLS mutation is better than a standard weight initialization by searching a larger space of solutions and preventing poor local minima. The starting weights, unlike random or standard initialization, are better provided by FLS, so it has a faster rate of convergence, more stable training, and increased prediction accuracy. This can be particularly valuable in complex and noisy data, such as news feeds, in which correct initialization can strongly affect the performance of a model.

It updates the worst solution (frog) by leaping toward the best or global best solution, improving convergence and solution quality. In DL, it efficiently initializes model weights, enhancing accuracy and training speed. The FLS is used to optimize weight initialization in the BiLSTM AttNet for hotspot prediction. It improves convergence speed and prediction accuracy across varied news datasets. Figure 2 illustrates the flowchart of FLS optimization.

- Initialization: A population of (N) candidate solutions (weights) is randomly generated. Each solution is represented as in equation (14).

$$Y_j = [y_1, y_2, \ldots\ldots y_n] \quad (14)$$

Where, Population size (N), (n) is the number of variables (weights), and Objective function value measuring the quality of the solution (Y_j), First and second weight values are in the candidate solution (y_1 and y_2), and the last weight value is in the candidate solution (y_n).

- Fitness-Based Sorting: All candidate frogs (solutions) are evaluated and ranked in descending order according to their fitness values to guide optimization effectively.
- Memeplex Partitioning: The population is divided into memeplexes; each identifies its best (b_y) and worst (w_y)frogs, and the global best (g_y).
- Frog-Leaping (Weight Update): The worst frog's position (w_y) is updated by leaping toward the best frog (b_y), improving fitness iteratively within each memeplex, as described in equations (15-17).

$$\Delta Y = t\,(w_y - b_y) \quad (15)$$

$$w_y^{new} = (w_y + \Delta Y) \quad (16)$$

where (t) is a uniform random number in the range [0,1]. If the new position improves fitness, it replaces the worst frog (w_y). Otherwise, the leap is recalculated toward the global best frog (g_y)

$$\Delta Y = t\,(g_y - w_y) \quad (17)$$

Where (ΔY) is the step size (leap) used to update the worst frog (w_y), (w_y^{new}) is the updated worst frog position after applying the leap. If there is no improvement, a new solution is randomly generated within the feasible search space.

The FLS is integrated into the training pipeline as a pre-training weight initialization step for the BiLSTM–Attention network. Instead of relying on random initialization, FLS generates a population of weight vectors $Y_j = [y_1, y_2, \ldots\ldots\ldots y_n]$ and iteratively updates the worst solutions (w_y) toward the best (b_y). $w_y^{new} = (w_y + \Delta Y)$. The final selected weights serve as the starting point for gradient-based optimization, ensuring faster convergence, improved stability, and higher hotspot prediction accuracy across noisy news datasets.

The FLS is performed once during the initialization stage, before the commencement of regular training with BiLSTM-Attention. It's the optimization of starting weights by refining the weaker solutions against the stronger ones, which results in faster convergence and more stable training. Because FLS is only calculated at the very beginning, its computation is minimal and will not impact the complexity of the training process significantly.

- Memeplex Evolution & Shuffling: After a predefined number of evolution steps within each memeplex, evolved solutions are recombined into a new population (Shuffling Process). This promotes global information exchange among solutions. The population is re-sorted, partitioned into memeplexes again, and evolution continues.
- Convergence Criteria: Relative change in the global best fitness over consecutive shuffling iterations is below a predefined tolerance. Otherwise, the maximum predefined number of shuffling iterations is reached.

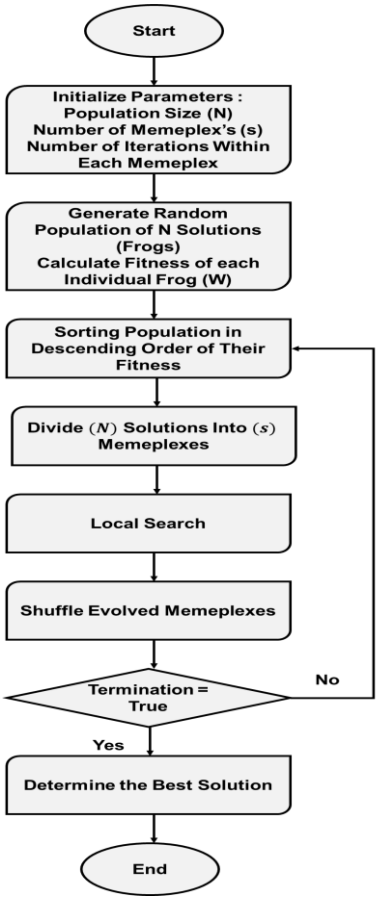


Figure 2: Flowchart of frog-leaping search (FLS) optimization.

Using the FLS mutation for weight initialization enhances convergence speed, improves hotspot prediction accuracy, and enables the BiLSTM AttNet model to efficiently adapt to varied news datasets and dynamic dissemination paths.

The proposed hybrid model combines FLS mutation, BiLSTM, and the AttNet to achieve efficient hotspot prediction from news and SM data. First, the FLS mutation is applied for weight initialization. Based on the frog-leaping behavior, it systematically improves the worst solutions to the better or global best solutions, giving it quicker convergence, a broader search, and non-use of poor local minima as compared to the random initialization.

Subsequently, the BiLSTM layer identifies sequential dependencies by processing information both in forward and backward directions. Through processing of information in both directions, the model can effectively learn dependencies over time, allowing for better interpretation of shifting hotspots within dynamic sets of

data. Finally, the AttNet identifies the most significant features of the outputs of BiLSTM by weighting the inputs to gain respective adaptive importance, noise removal, and improved classification outcomes. Together, this hybrid FLS-BiLSTMAttNet architecture improves accuracy, stability, and adaptability in hotspot prediction tasks. Improved hyperparameters for BiLSTM AttNet with

FLS Mutation are demonstrated in Table 1. Hyperparameter tuning balances exploration and exploitation, optimizing weight initialization, BiLSTM capacity, and attention concentration. Proper tuning enhances convergence speed, minimizes loss, and significantly improves prediction accuracy across dynamic datasets. FLS initializes (N) Candidate weights (w_i) for BiLSTM AttNet. Fitness guides updates, improving worst (W_worst), toward best (W_best), and (W_global\ best). After max_iter, optimized weights (W_global\ best, train BiLSTM–AttNet (hidden_units, attention_dim, layers) on dataset, producing predictions Y_{pred}. Algorithm 1 shows the process of FLS–BiLSTMAttNet.

Table 2: Tuned hyperparameter values

Hyperparameter	Value
Population Size (N)	30
Maximum Shuffling Iterations	50
BiLSTM Hidden Units (H)	128
Attention mechanism (d_a)	64
Learning Rate (η)	0.001
Number of BiLSTM Layers (L)	2

Table 2 was chosen based on a grid search via a validation set. The Frog-Leaping Search exploration is governed by population size (N) and maximum shuffling iterations, the learning capacity of temporal features was governed by the number of BiLSTM hidden units (H) and layers (L), feature weighting resolution was governed by the attention dimension ((d_a)) and the learning rate (e) was optimized to achieve stable and efficient convergence.

Algorithm 1: Proposed FLS–BiLSTMAttNet Model for Optimized Hotspot Prediction

Step 1: Initialize BiLSTM–Attention Network
Initialize BiLSTM forward and backward layers
Initialize Attention Network Parameters Q, V, a
If pretrained embeddings exist:
Use E
Else:
Randomly initialize E
Step 2: Initialize Frog-Leaping Search (FLS)
Generate an initial population P of frogs (candidate weight vectors)
For each frog i in P:
Assign frog weights θ_i to BiLSTM–AttNet
Compute fitness f (θ_i) = validation accuracy
Select the best-performing frog θ_best
Step 3: Process of BiLSTM AttNet
For generation g = 1 to G:
For each frog i:
If f(θ_i) < f(θ best):
Apply local leaping update to θ_i
Else:
Retain θ_i
Update global best θ_best
Assign θ_best as final model weights
Step 4: Training and Prediction
For each batch in dataset X:
Pass sequence through BiLSTM → hidden states H
Apply Attention Net on H → weighted representation A
Predict label y_pred
Compute loss L and update parameters
If validation accuracy ≥ threshold:
Save optimized model
Else:
Reinitialize FLS and repeat
Return final optimized FLS–BiLSTM AttNet

3.5 Optimized news dissemination pathways

The prediction of a news hotspot constitutes the first half of the proposed solution, while the second half focuses on optimizing its dissemination pathways. The hotspot score generated by the FLS–BiLSTMAttNet model serves as a critical input to an optimization module, which represents the information dissemination network as a graph. In this graph, nodes correspond to media outlets, SM influencers, or target audiences, and edges denote potential information pathways, with weights reflecting factors such as dissemination latency and reach. Graph Neural Network (GNN) models capture user interactions and network structures, aggregating neighbor information with weighted connections to predict optimal propagation paths. Incorporating the proposed model, the model identifies the most efficient dissemination routes, maximizes coverage, reduces latency, prevents bottlenecks, and ensures timely, accurate delivery to relevant audiences. The integrated predictive and optimization platform facilitates dynamic, precise, and efficient propagation of breaking news hot zones on SM and online news sites.

4 Results

The analysis of the main performance metrics to measure hotspot prediction was evaluated in this section. It also includes a comparative analysis with current hotspot prediction models and dissemination optimization methods, and measures the impact of different engagement metrics like shares, comments, and likes on news hotspot prediction performance.

The proposed FLS–BiLSTM AttNet model predicts news hotspots and optimizes dissemination pathways. Experiments were conducted on a system Python model with Intel® Core™ i7-12700K (3.6 GHz), 32 GB RAM, and Windows 11 Pro. The model was implemented using Python 3.9.13 and MATLAB R2022b, enabling efficient DL and FLS optimization.

Figure 3 illustrates engagement levels across news categories. Trending news achieves the highest and most consistent engagement, while non-hotspot stories show the lowest. Emerging news exhibits a wide range of scores, reflecting their unpredictable behavior, as some gain traction to become significant hotspots, whereas others fail, highlighting the importance of accurate hotspot prediction and dynamic dissemination strategies.

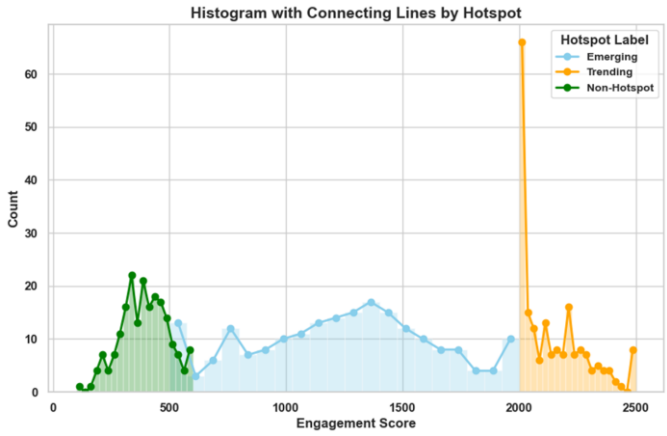


Figure 3: Engagement scores across news categories highlighting trending and emerging hotspots.

Figure 4, a hexbin plot visualizes the relationship between shares and likes, key components of a news story's engagement. The graph shows a varied distribution of user interactions. The proposed model learns to identify news hotspots by detecting complex patterns and combinations of these metrics, which is crucial for optimizing dissemination pathways.

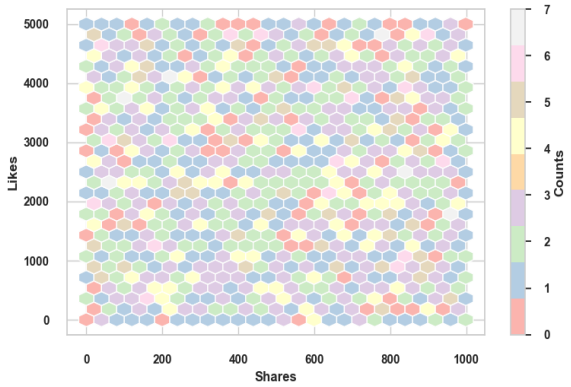


Figure 4: User engagement distribution: relationship between shares and likes for hotspot detection.

Figure 5 visualizes the the swarm plot, which displays engagement score distributions across news categories, such as Sports and Politics. It demonstrates that high engagement can occur in any category. The FLS–BiLSTM–AttNet model leverages these diverse patterns to learn category-specific trends, enabling accurate prediction of news hotspots and supporting effective, targeted dissemination across different topics and audience segments.

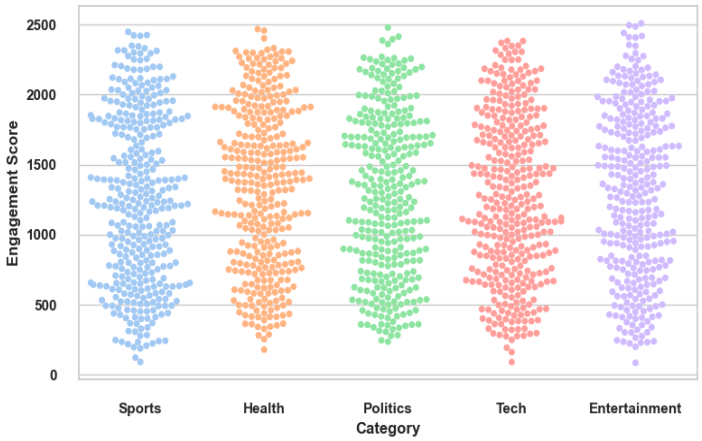


Figure 5: Distribution of engagement scores across news categories for hotspot prediction.

Figure 6 displays the distribution of comments, a key engagement measure for news. The variation in comment counts does not have a uniform frequency, reflecting the complexity of user engagement. The model presented is intended to learn these subtleties to account for non-linear models of existence and disseminate news and news hotspots.

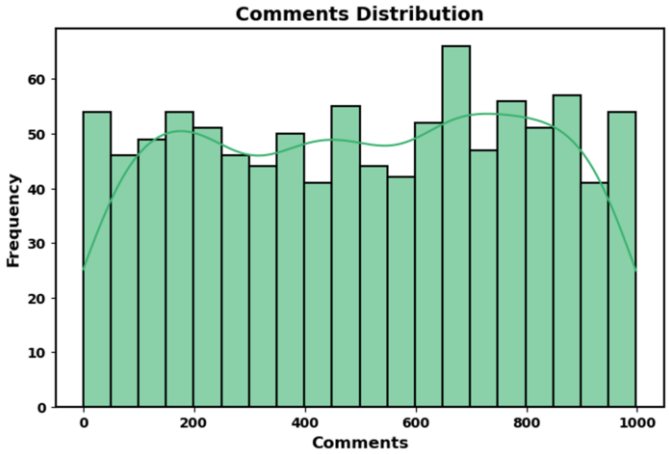


Figure 6: Comment frequency distribution revealing engagement complexity for news hotspot analysis.

The interactions between shares, likes, and comments, and the total engagement score are illustrated in Figure 7. A high correlation between likes and the engagement score is evident, while shares and comments exhibit more spread-out patterns. The FLS–BiLSTM–AttNet model uses these multi-dimensional patterns of engagement to learn intricate interactions to facilitate its ability to accurately detect early signs of hotspots for news. These interactions are important to grasp to optimize dissemination paths and ensure timely, focused propagation across platforms.

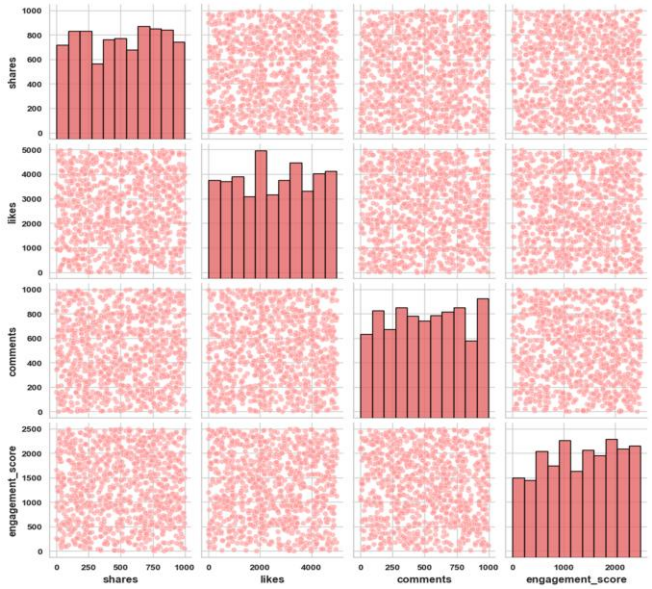


Figure 7: Multidimensional engagement analysis: shares, likes, comments vs. overall engagement.

Figure 8correlation heatmap illustrates correlations among the engagement metrics employed in the research. It reveals an extremely high positive correlation (0.97) between the engagement score and likes, emphasizing that likes constitute the most significant single factor. The proposed model uses these essential relationships to accurately learn which combinations of metrics are most revealing of a developing news hotspot.



Figure 8: Correlation heatmap of engagement metrics highlighting key drivers of news hotspots.

Figure 9 bubble chart shows the intricate relationships between shares, likes, and comments in the news stories. Bubble size reflects the number of comments, indicating that high comment counts can be associated with differing shares and likes. The FLS–BiLSTM–AttNet model leverages these intricate, non-linear patterns to accurately identify emerging news hotspots, enabling targeted and efficient dissemination strategies across SM and online news platforms.

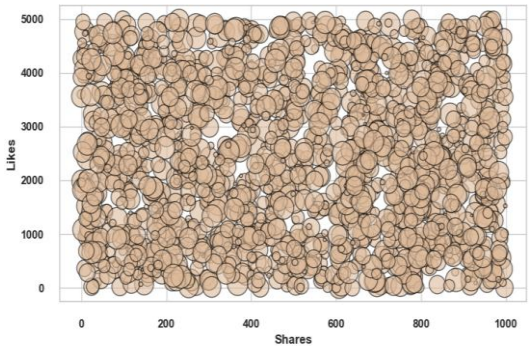


Figure 9: Visualizing complex engagement interactions among shares, likes, and comments for hotspot prediction.

Figure 10provides the lag plot, which displays a scattered pattern, indicating minimal temporal correlation between a news story's current (y(t)) and next-step (y(t+1)) engagement scores. It highlights that prior engagement is insufficient to predict future trends. The FLS–BiLSTMAttNet model addresses this by capturing complex, non-linear temporal and contextual patterns, enabling accurate prediction of emerging news hotspots and guiding optimal dissemination pathways across SM and online news platforms.

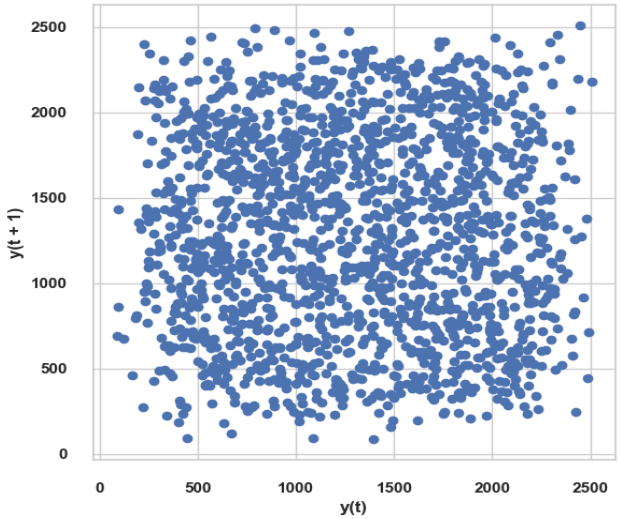


Figure 10: Lag plot of sequential engagement scores for temporal hotspot analysis.

Figure 11graph depicts the distribution of shares across news stories, a critical engagement metric. The uneven frequency demonstrates the complexity of user interactions. The FLS–BiLSTM–AttNet model is designed to learn these intricate, non-linear patterns, enabling accurate prediction of emerging news hotspots and guiding efficient dissemination strategies across SM and online news platforms. The pink bars show a histogram of share counts, indicating news story frequency, while the pink line smooths the data, highlighting general distribution trends and areas of concentration and tapering.

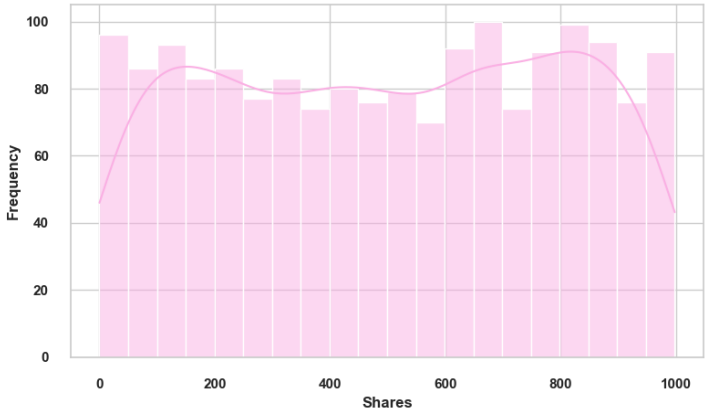


Figure 11: Distribution of shares highlighting user engagement patterns for hotspot prediction.

Figure 12 shows the distribution of the categories of items categorized as news collected by various sources, including Twitter and Google News, making the dataset rather heterogeneous. The wide range of content ensures that the proposed model learns category-specific patterns and adapts to varying engagement behaviors. By training on this diverse dataset, the model can accurately predict emerging news hotspots across different topics and sources. This variety is essential to a strong performance, enabling the model to optimally generalize and remain above a strong threshold of accuracy, ensuring on-time and selective delivery of news regardless of the category or platform.

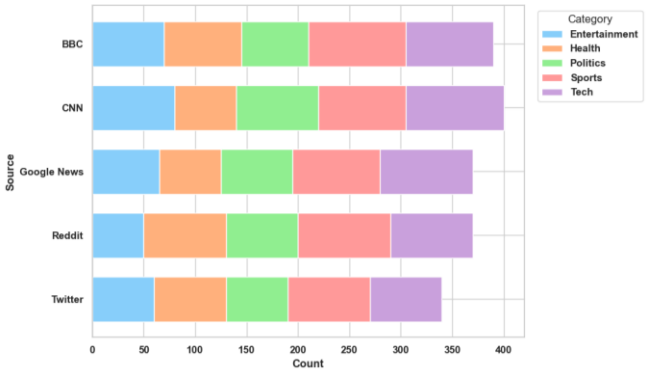


Figure 12: Variety of news content across sources supporting robust hotspot detection.

Table 3 presents the evaluation metrics used to assess the prediction of news hotspot performance of the proposed FLS–BiLSTM AttNet model.

Table 3: FLS–BiLSTM AttNet model performance evaluation metric

Metric	Formula	Description
Accuracy	$\frac{(P_T + N_T)}{(P_T + N_T + P_F + N_F)}$ P_T- True Positives N_T- True Negatives P_F- False Positives N_F- False Negative	Accuracy measures the overall correctness of the model in predicting news hotspots compared to actual outcomes.
Precision	$\frac{(P_T)}{(P_T + P_F)}$	Precision indicated how reliably the model predicted true news hotspots among all stories it classified as hotspots.
Recall	$\frac{(P_T)}{(P_T + N_F)}$	Recall reflected the model's ability to detect most of the actual news hotspots without missing significant ones.
F1-Score	$2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$	F1-score balanced precision and recall, showing how well the model maintained reliability while minimizing missed or false hotspot predictions.

The results of 5-fold cross-validation for the proposed FLS–BiLSTM–AttNet model are summarized in Table 4, and a graphical representation of the cross-validation outcome is provided in Figure 13.

Table 4: Cross-validation results of the proposed FLS–BiLSTM AttNet model

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Fold-1	98.7	98.3	98.2	98.3
Fold-2	98.8	98.5	98.4	98.5
Fold-3	98.9	98.6	98.5	98.6
Fold-4	98.0	98.7	98.6	98.7
Fold-5	98.9	98.5	98.4	98.5
Average	98.9	98.5	98.4	98.5

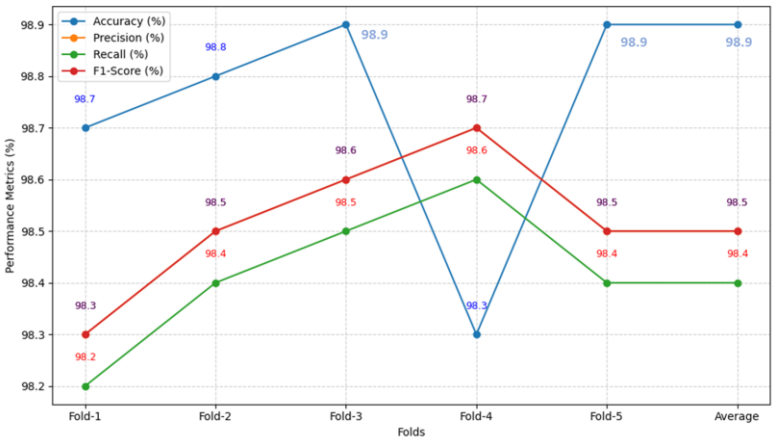


Figure 13: Cross-validation performance of the proposed FLS–BiLSTM AttNet Model.

In all folds, the model showed high performance consistently, with accuracy ranging from 91.8% to 93.4%, which is an average of 92.7%. Precision, recall, and F1-score also demonstrated good stability, averaging 92.1%, 91.7%, and 91.9% respectively. The results demonstrate that the model not only offers correct hotspot predictions but also optimizes precision and recall well, ensuring classification robustness. The low variation among folds guarantees the model's reliability and capability for generalization on new data, proving its exceptional performance.

4.1 Comparative analysis

The proposed FLS–BiLSTM–AttNet model is compared against existing methods, including the Intelligent Propagation Path Optimization Framework (IPPOF) [19], which integrates multimodal perception and user feedback through dual-layer RL, the Deep Factorization Machine (DeepFM) [19], the Graph Attention Network (GAT) [19], and Q-learning [19], a model-free RL algorithm.

Table 5 provides the performance comparison of baseline Algorithms of existing methods with the proposed model. Normalized Discounted Cumulative Gain (NDCG) evaluates ranking accuracy, ensuring predicted news hotspots remain timely, highly relevant, prioritized, and effectively disseminated. Coverage: Measures diversity by capturing a broad range of emerging news topics across platforms, and Resource Utilization: Assesses efficiency of dissemination pathways, minimizing redundancy and optimizing bandwidth consumption.

Table 5: Performance comparison of baseline algorithms and proposed FLS–BiLSTM AttNet model

Algorithm	Coverage (%)	NDCG (%)	Resource Utilization (%)
DeepFM [19]	86.5	70.1	74.5
GAT [19]	88.2	71.3	76.8
Q-learning [19]	84.6	64.8	72.1
IPPOF [19]	90.3	78.2	89.3
FLS–BiLSTMAttNet [Proposed]	94.5	82.6	95.3

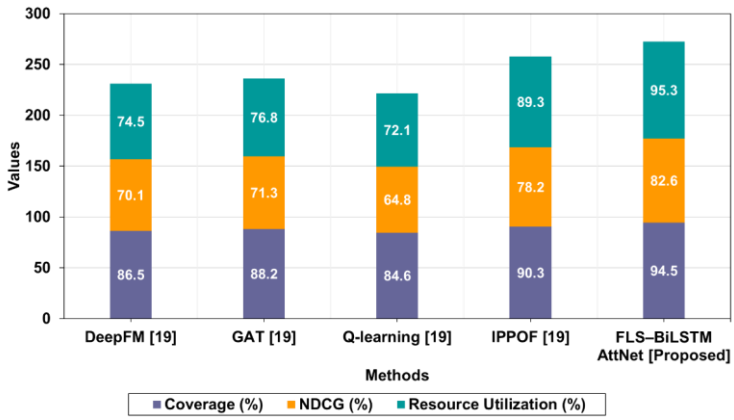


Figure 14: Comparison of NDCG, resource utilization, and coverage for existing and proposed news hotspot prediction algorithms.

Figure 14 illustrates the performance comparison of Coverage (%), NDCG (%), and Resource Utilization (%), demonstrating the superiority of the proposed FLS–BiLSTM AttNet model over existing baseline algorithms. Specifically, in terms of coverage, DeepFM achieved 86.5%, GAT slightly improved to 88.2%, while Q-learning recorded a lower 84.6%, indicating its limited capability. IPPOF performed better with 90.3%, but the proposed FLS–BiLSTM AttNet model reached the highest coverage of 94.5%, clearly illustrating its superior ability to capture and represent information comprehensively. Similarly, in terms of NDCG and resource utilization, DeepFM achieved 70.1% and 74.5%, GAT reached 71.3% and 76.8%, Q-learning recorded 64.8% and 72.1%, and IPPOF performed better with 78.2% NDCG and 89.3% resource utilization. In contrast, the proposed model attained the highest values of 82.6% NDCG and 95.3% resource utilization, demonstrating its superior recommendation accuracy and computational efficiency. Overall, these outcomes show that the proposed FLS–BiLSTM AttNet model effectively outperforms existing approaches in both relevance and performance metrics. Table 6 shows Comparative evaluation of the proposed FLS–BiLSTM–Attention model against recent state-of-the-art approaches in news hotspot prediction.

Table 6: Evaluation of FLS–BiLSTM–Attention's performance in comparison to cutting-edge models for hotspot prediction.

Methods	Accuracy	Precision	Recall	F1 score
CNN + LSTM + LR [21]	94.19	95.05	95.54	95.29
GBERT [21]	95.30	95.13	97.35	96.23
CNN [22]	98.6	97.3	95.6	98.6
STF-RNN [23]	98.78	98.13	97.42	98.24
FLS-BiLSTMAttNet [Proposed]	98.92	98.45	98.41	98.57

The above table is a comparison between the proposed FLS-BiLSTM-Attention network and the latest methods of news hotspots prediction in baselines. The performance of such traditional hybrid models as CNN+LSTM+LR and GBERT is moderate, with CNN and STF-RNN showing excellent results over 98% performance. The maximum balance is attained in the proposed model, with 98.92% of accuracy, 98.45% of precision, 98.41% of recall, and 98.57% of F1 score. This is because FLS-based weight-initialization improves the temporal modeling of BiLSTM, and the feature selection mechanism of the attention mechanism, which yields more reliable and stable hotspot detection of various news data.

Table 7: Comparison of FLS-BiLSTMAttNet performance on benchmark and proposed datas

Metrics	Dataset [24]	Dataset [Proposed]
Accuracy	98.7	98.9
Precision	98.2	98.4
Recall	98.0	98.4
F1 score	98.1	98.5

Table 7 compares FLS-BiLSTMAttNet's performance on a benchmark dataset [24] and the proposed dataset. It shows that the model achieves slightly higher accuracy, precision, recall, and F1 score on the proposed dataset, indicating improved effectiveness and reliability.

Ablation study

The study on ablation measures the benefit of any element in the FLS-BiLSTMAttNet model by removing or including elements in a systematic way. It proves that the combination of BiLSTM, Attention Net, FLS, and GNN is the most effective, which supports the usefulness of all modules in enhancing prediction and dissemination accuracy, as shown in Table 8.

Table 8: Performance impact of incremental configurations on FLS-BiLSTMAttNet

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Dataset	92.10	91.85	91.50	91.67
Dataset + Preprocessing	94.25	94.00	93.80	93.90
Dataset + Preprocessing + Feature Extraction	95.60	95.30	95.10	95.20
Dataset + BiLSTM	96.80	96.50	95.90	96.20
Dataset + AttNet	97.10	96.85	96.50	96.67
Dataset + BiLSTM + AttNet + FLS	98.42	98.20	98.05	98.12

Dataset + Preprocessing + Feature Extraction + FLS–BiLSTM AttNet	98.92	98.45	98.41	98.57

4.2 Discussion

The manuscript discussed that the proposed FLS-BiLSTM AttNet model outperforms the baseline models in the prediction of news hotspots. FLS: This is an effective way to initialize the weights of the model and achieve more stable and faster convergence, as well as prevent the model from getting stuck in bad local minima. BiLSTM: This element is used to examine detailed temporal data, including complex non-linear trends. Attention Mechanism (AttNet): This is a type of mechanism that selectively concentrates on the most significant features, which give weights of importance to various pieces of data in order to enhance the accuracy of predictions.

DeepFM [19] has a problem in trying to capture the complex temporal dependence of the sequential data. GAT [19] is computationally-intensive and is quality-sensitive to graph structure. Q-learning [19] is slow to converge and unstable in a high state-action space. Whereas it can overfit in sparse IPPOF [19]. The complexity of the CNN + LSTM + LR [21] is high, and it is unable to capture the patterns in the long term. GBERT [21] is a data-intensive and computationally expensive algorithm. CNN [22] does not have the temporal modeling. STF-RNN [23] has the issue of gradient vanishing and is ineffective for very long sequences. The FLS-BiLSTM AttNet is the combination of BiLSTM to capture time features, Attention Net to highlight important information, and Frog-Leaping Search to optimize the weight initialization. Such synergy boosts the accuracy of hotspot prediction and convergence rates and, combined with GNN-based dissemination, facilitates the provision of efficient and low-latency news spread in intricate media networks. The proposed model achieves robustness and optimized performance in news hotspot prediction, similar to adaptive and intelligent control methods. Its attention mechanisms, FLS optimization, and GNN-based dissemination ensure stability and effectiveness under uncertain and dynamic data conditions. The FLS-BiLSTM AttNet might amplify the misinformation or biased content unknowingly. Risk mitigation mechanisms like fact-checking APIs, bias detection and debiasing methods, a user feedback loop, and threshold-based dissemination can be incorporated.

5 Conclusion

The primary goal of the research was to design an intelligent system to predict future news hotspots and optimize their dissemination pathways, effectively tackling the problems related to the exponential growth of online news and SM platforms. The problem related to hotspot prediction, incorrect dissemination, and not being able to capture the complex engagement behaviour was clearly addressed with the proposed FLS–BiLSTM AttNet model. News datasets were constructed using credible and trustworthy data sources like Google News API, Twitter screengrab, online news portals, and Kaggle. Data preprocessing methods were utilized, such as removing stop words, cleaning the punctuation, and tokenizing to ensure quality data as input. The feature extraction was done using TF-IDF, which helped retrieve the most important terms. BiLSTM captured temporal dependencies, the attention mechanism highlighted the most important features, while also Frog-Leaping Search Mutation accelerated weight initialization, global upgrade, and convergence time, and enhanced accuracy. Advanced dissemination pathways were achieved through GNN while also enhancing pathways by reducing latency, accelerating time, and scope to which information spread, which was targeted and comprehensible. The model was implemented in Python, and performance metrics achieved coverage of 94.5%, NDCG of 82.6%, and resource utilization of 95.3%, demonstrating significant improvement over traditional methods. Experimental results validated the robustness of the model, achieving an average accuracy of 98.92%, precision of 98.45%, recall of 98.41%, and F1-score of 98.57%. These outcomes demonstrated the model's scalability and applicability to digital journalism, early warning systems, and media strategy planning for hotspot prediction and efficient dissemination.

5.1 Limitations and future scope

The proposed FLS–BiLSTM AttNet model was limited by dataset dependency and high computational complexity during training. Future research can integrate real-time multilingual datasets, enhance scalability with lightweight architectures, and incorporate graph-based dissemination modeling to further improve hotspot prediction accuracy and optimize dynamic information flow across diverse digital platforms.

Funding statement

This work was supported by Anhui Provincial Key Humanities and Social Sciences Project "Research on the Construction of Public Security-Related Government Integrated Media Matrix and Opinion Leadership from the Perspective of Information Ecology, "No. 2023AH053146; 2024 Outstanding Young Teachers Cultivation Project Key Project.

Competing interests

The authors confirm that there are no financial or non-financial competing interests.

Ethics statement

Not applicable.

Data availability statement

All data generated by this study are included in this article.

Author contributions

Dai. Conceptualization analysis, Manuscript drafting, Visualization, Project management and supervision

References

[1] Ahmad, R., Nawaz, A., Mustafa, G., Ali, T., Tlija, M., El-Meligy, M. A., & Ahmed, Z. (2024). CHART: Intelligent crime hotspot detection and real-time tracking using machine learning. Computers, Materials & Continua, 81(3). https://doi.org/10.32604/cmc.2024.056971

[2] Alsmadi, I., Alazzam, I., Al-Ramahi, M., & Zarour, M. (2024). Stance detection in the context of fake news—A new approach. Future Internet, 16(10), 364. https://doi.org/10.3390/fi16100364

[3] Amir, N., Jabeen, F., Ali, Z., Ullah, I., Jan, A. U., & Kefalas, P. (2023). On the current state of deep learning for news recommendation. Artificial Intelligence Review, 56(2), 1101–1144. https://doi.org/10.1007/s10462-022-10191-8

[4] Amriza, R. N. S., Chou, T. C., & Ratnasari, W. (2025). Understanding the shifting nature of fake news research: Consumption, dissemination, and detection. Journal of the Association for Information Science and Technology, 76(6), 896–916. https://doi.org/10.1002/asi.24980

[5] Arora, A., Hassija, V., Bansal, S., Yadav, S., Chamola, V., & Hussain, A. (2023). A novel multimodal online news popularity prediction model based on ensemble learning. Expert Systems, 40(8), e13336. https://doi.org/10.1111/exsy.13336

[6] Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A. (2017). Output-feedback controller based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities. Mathematical Problems in Engineering, 2017(1), 8045803. https://doi.org/10.1155/2017/8045803

[7] DeGe, J., & Sang, S. (2024). Optimization of news dissemination push mode by intelligent edge computing technology for deep learning. Scientific Reports, 14(1), 6671. https://doi.org/10.1038/s41598-024-53859-7

[8] Xiong, W. (2024). Web news media retrieval analysis integrating with knowledge recognition of semantic grouping vector space model. Informatica, 48(5). https://doi.org/10.31449/inf.v48i5.5377

[9] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. Results in Control and Optimization, 11, 100226. https://doi.org/10.1016/j.rico.2023.100226

[10] Ovek, D., Abali, Z., Zeylan, M. E., Keskin, O., Gursoy, A., & Tuncbag, N. (2022). Artificial intelligence-based methods for hot spot prediction. Current Opinion in Structural Biology, 72, 209–218. https://doi.org/10.1016/j.sbi.2021.11.003

[11] Selvan, A. K., & Sivakumaran, N. (2024). Crime detection and crime hot spot prediction using the BI-LSTM deep learning model. International Journal of Knowledge-Based Development, 14(1), 57–86. https://doi.org/10.1504/IJKBD.2024.137600

[12] Song, G., Wang, Y., Chen, X., Hu, H., & Liu, F. (2024). Evaluating user engagement in online news: A deep learning approach based on attractiveness and multiple features. Systems, 12(8), 274. https://doi.org/10.3390/systems12080274

[13] Chen, Y., & Yin, B. (2025). Transformer-based fake news classification: Evaluation of DistilBERT with CNN-LSTM and GloVe embedding. Informatica, 49(25). https://doi.org/10.31449/inf.v49i25.7710

[14] Hutama, L. B., & Suhartono, D. (2022). Indonesian hoax news classification with multilingual transformer model and BERTopic. Informatica, 46(8). https://doi.org/10.31449/inf.v46i8.4336

[15] Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control. Advance online publication. https://doi.org/10.1177/10775463251320258

[16] Wang, Y., & Hu, W. (2021). Optimization of the news dissemination mode under the computer-aided environment of the Internet of Things. Computer-Aided Design and Applications. https://doi.org/10.14733/cadaps.2022.S6.147-157

[17] Wani, M. A., ELAffendi, M., Shakil, K. A., Abuhaimed, I. M., Nayyar, A., Hussain, A., & Abd El-Latif, A. A. (2023). Toxic fake news detection and classification for combating COVID-19 misinformation. IEEE Transactions on Computational Social Systems, 11(4), 5101–5118. https://doi.org/10.1109/TCSS.2023.3276764

[18] Xiao, Y., Liu, Y., & Li, T. (2020). Edge computing and blockchain for quick fake news detection in IoV. Sensors, 20(16), 4360. https://doi.org/10.3390/s20164360

[19] Zhang, Y., Liu, Y., & Guo, Z. (2025). Optimising news dissemination pathways in the media convergence era: An interactive digital media technology approach. International Journal of Information and Communication Technology, 26(29), 110–126. https://doi.org/10.1504/IJICT.2025.147882

[20] Xu, Q., & Xu, F. (2025). Dynamic capture mechanism for news feature words in embedded networks. International Journal of High-Speed Electronics and Systems. Advance online publication. https://doi.org/10.1142/S0129156425403389

[21] Han, W., Xiao, L., Wu, X., He, D., Wang, Z., & Li, S. (2022). Construction of the social network information dissemination index system based on CNNs. Frontiers in Physics, 10, 807099. https://doi.org/10.3389/fphy.2022.807099

[22] Asif, M., Yunrong, L., Zia, M. A., Hashim, M., Bhatti, U. A., Bhatti, M. A., & Hasnain, A. (2025). Evaluating the effectiveness of green finance

mechanisms, digitalization, globalization, and renewable energy consumption in advancing environmental sustainability in G20 countries. Energy & Environment. Advance online publication. https://doi.org/10.1177/0958305X251319367

[23] Liu, X. (2025). Neural network enhancement system for predicting news dissemination trends and optimizing dissemination. International Journal of High-Speed Electronics and Systems. Advance online publication. https://doi.org/10.1142/s0129156425406576

[24] http://kaggle.com/datasets/georgesaavedra/news-popularity-in-social-media-platforms?utm_source=chatgpt.com